ATARNotes

General Maths 3/4

Unit 3 Head Start

January Lecture Series

Presented by:

Josh Hamilton

Welcome!

ATARNotes

HIGH SCHOOL RESOURCES

Success for all students

atarnotes.com/register

HEAPS MORE

ATAR Notes FREE resources

STUDY NOTES
Hundreds downloadable

LECTURES
Learn from the best

DISCUSSIONS
Online Q&A

VIDEOS
Engaging online revision

NEWSLETTERS
Stay in the loop

ATAR CALC
See if you're on track

ARTICLES
Study strategy & tips

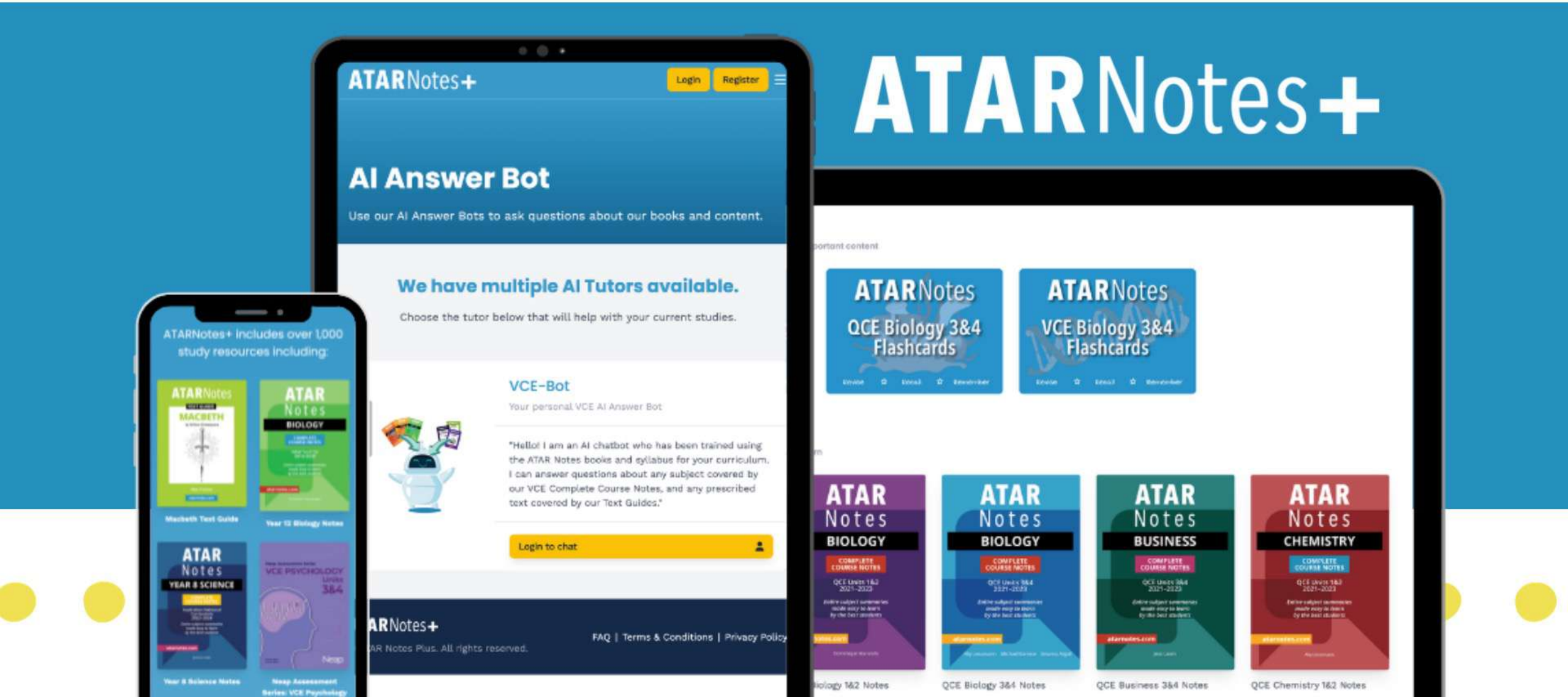
HEAPS MORE
Check out our website

atarnotes.com/register

HEAPS MORE

ATARNotes

Thousands of Study Resources



Welcome!!

Topics to be covered

- Welcome to the General Maths 3/4 head start lecture for 2024 !

House keeping:

- Please feel free to utilise the chat to ask any questions
- The slides should be able to be accessed below
- The recording will be available after the lecture premier

Lecture Outline

20 minutes

BLOCK 1: OVERVIEW & STUDY TIPS

40 minutes

BLOCK 2: UNIVARIATE DATA

60 minutes

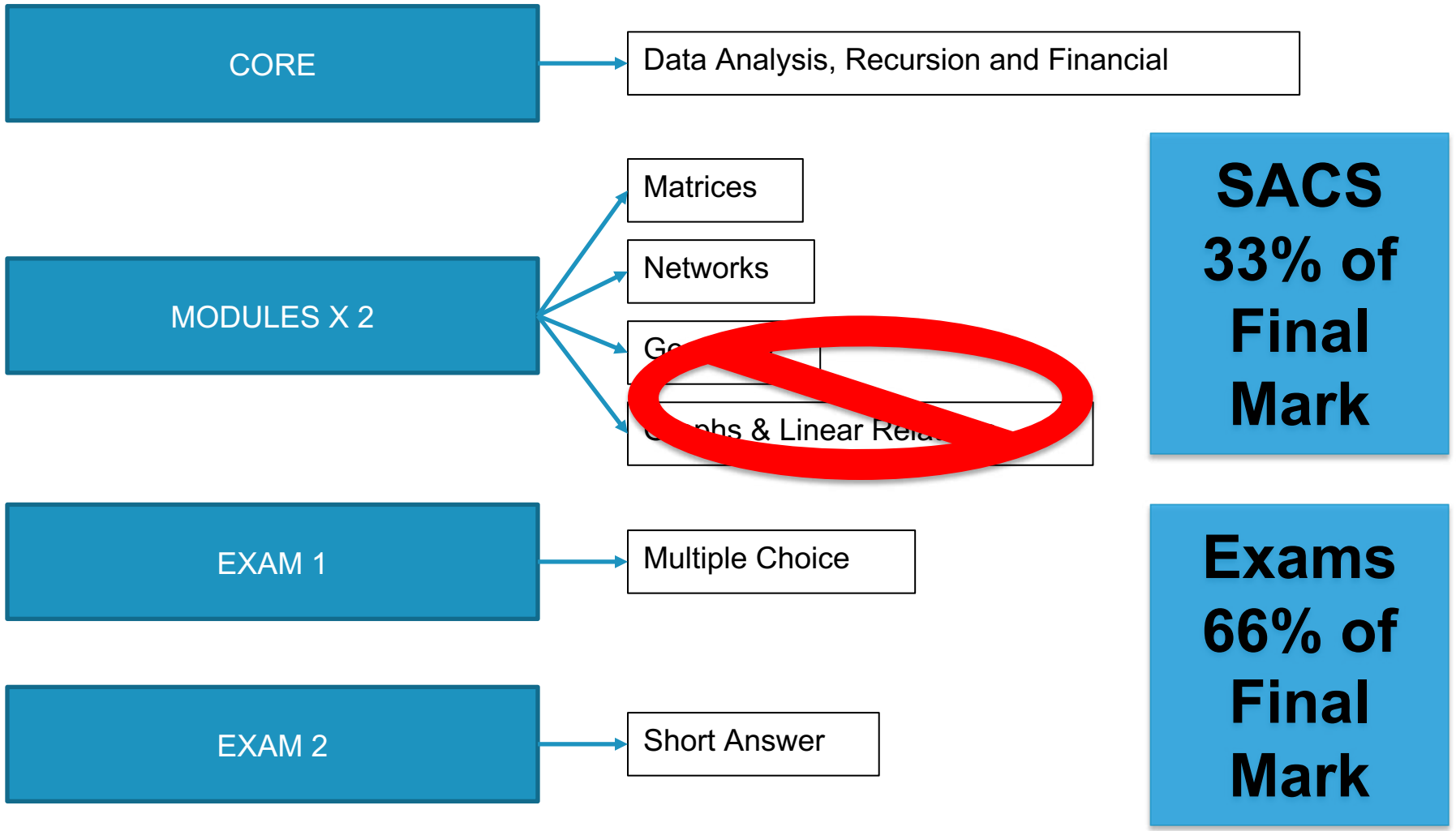
BLOCK 3: BIVARIATE AND APPLICATION OF DATA

Who am I?

ATARNotes

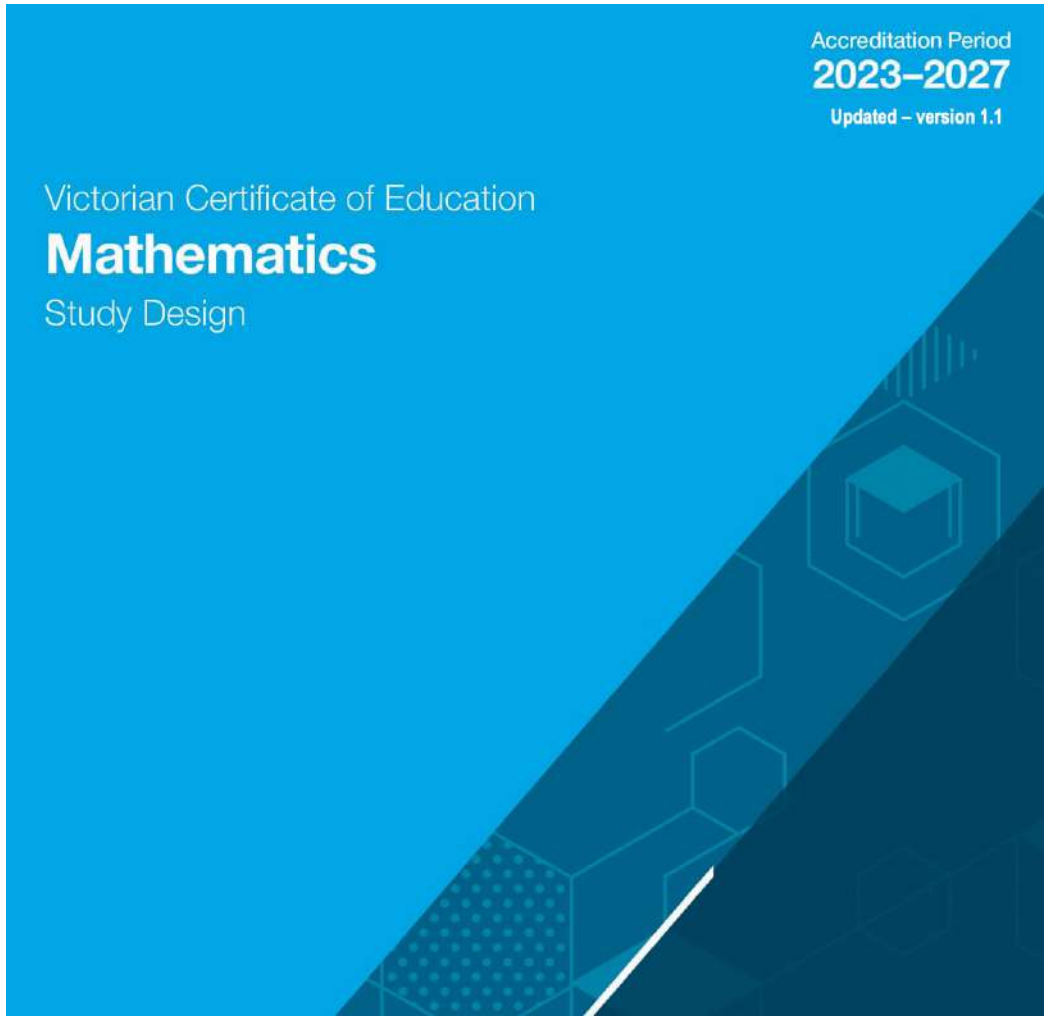
Course Overview and Tips

General Math Structure



Overview & Tips

Study Design Change (2023)



Read the study design!
It changed in 2023!

Units 3 and 4: General Mathematics

General Mathematics Units 3 and 4 focus on real-life application of mathematics and consist of the areas of study 'Data analysis, probability and statistics' and 'Discrete mathematics'.

Unit 3 comprises *Data analysis* and *Recursion and financial modelling*, and Unit 4 comprises *Matrices* and *Networks and decision mathematics*.

Assumed knowledge and skills for General Mathematics Units 3 and 4 are contained in General Mathematics Units 1 and 2, and will be drawn on, as applicable, in the development of related content from the areas of study, and key knowledge and key skills for the outcomes of General Mathematics Units 3 and 4.

In undertaking these units, students are expected to be able to apply techniques, routines and processes involving rational and real arithmetic, sets, lists, tables and matrices, diagrams, networks, algorithms, algebraic manipulation, recurrence relations, equations and graphs. They should have facility with relevant mental and by-hand approaches to estimation and computation. The use of numerical, graphical, geometric, symbolic statistical and financial functionality of technology for teaching and learning mathematics, for working mathematically, and in related assessment, is to be incorporated throughout each unit as applicable.

Area of Study 1

Data analysis, probability and statistics

Data analysis

Students cover data types, representation and distribution of data, location, spread, association, correlation and causation, response and explanatory variables, linear regression, data transformation and goodness of fit, times series, seasonality, smoothing and prediction.

What changed for last year?

1. No more modules choice
   - This means that now schools do not get to choose 2 of the 4 smaller modules and must all do Networks + Matrices
2. Removal of non-casual effect and population statistics / sampling
   - This was low yield content in Data that has now been removed
   - Not many exam questions were asked on this
3. Removal of simultaneous equations and representation of linear lines in matrices
   - This was super high yield content that had many exam questions asked about it
   - Important to ignore these questions when doing practice questions
4. Addition of Leslie Matrix
   - We will cover this in a later lecture
5. Lastly, the change of name, Further → General

Overview & Tips

- Calculator guides in textbook – copy them out into your notes in case you go blank in a SAC or exam!
- Your CAS is your best friend, make sure you know it well
  - For my trigonometry friends, you WILL love your CAS cause bearing are a thing and can be a massive pain – so please learn how to use it well!
- Shortcuts through menu screens (e.g. menu – 3 – 1 for the solve function)
- 2 types of calculator, beware! – Ti nspire and the Casio Classpad, use the one your school makes you get!

Textbooks and Study Guide

- Complete all textbook exercises – even if your teacher says only to do left hand side or something like that – do ALL the questions!!!
  - Do these THROUGHOUT the year, NOT at the end!
  - The Textbook questions are amazing for first grasping and building a basic understanding of concepts – it is however NOT useful in building your application and examination skills

- Review sections – Identify your weak points in a specific topic
- Cut things out of your textbooks and study guides for your summary book if you find the topics helpful

- External study guides
  - ATAR Notes Course Guides/Topic Tests

- There are two (maybe three) models for a summary book that students follow:
  - Student A who puts literally everything he/she sees in their textbooks/study guides and class notes. At the end of the year their reference book is THICC but they have the peace of mind that everything's in there.
    - The negatives:
      - It will be a pain to look through during SACs or Exams and might waste valuable time.
      - Might consume a lot of your study time making it – if using this method, you NEED to stay up to date!
    - The positives:
      - You have the peace of mind that you have all the content and can use it when you get stuck.
      - Creating a comprehensive summary book is a good tool to revise content

- The second type of summary book:
  - Student B who would put together a minimal summary book with only the essential formulas – usually just a couple of pages printed off from a summary sheet found online.
    - The positives:
      - You spend more time doing practice questions – SUPER IMPORTANT!
      - Don’t waste as much time in an Exam/SAC flicking through your reference book
    - The negatives:
      - You're on your own in an Exam/SAC if you get stuck – all you have are formulas and only the key pieces of information.
- But the one golden rule is… you must create your OWN summary book!

ATARNotes

Area of Study 1

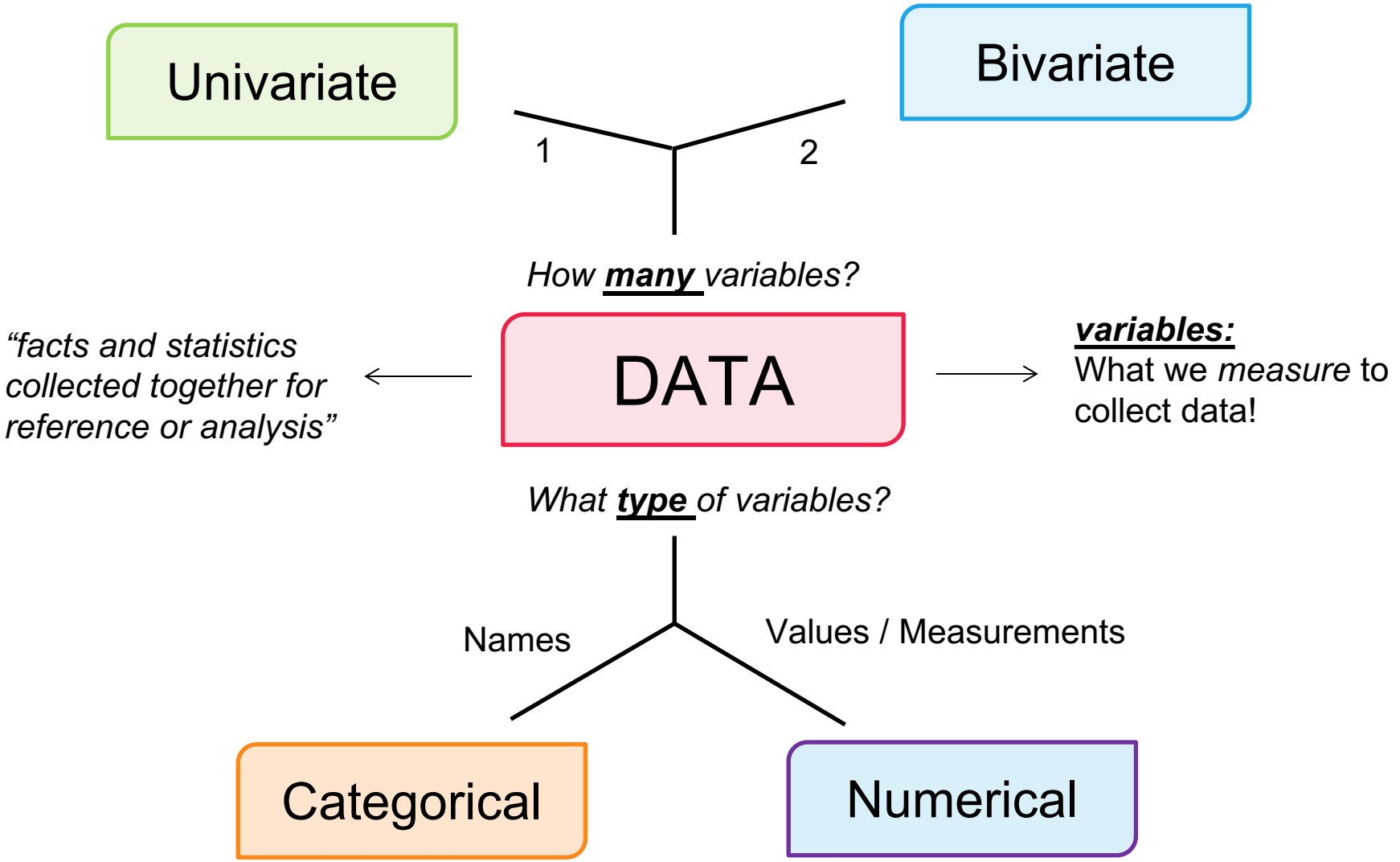
Data!

ATARNotes

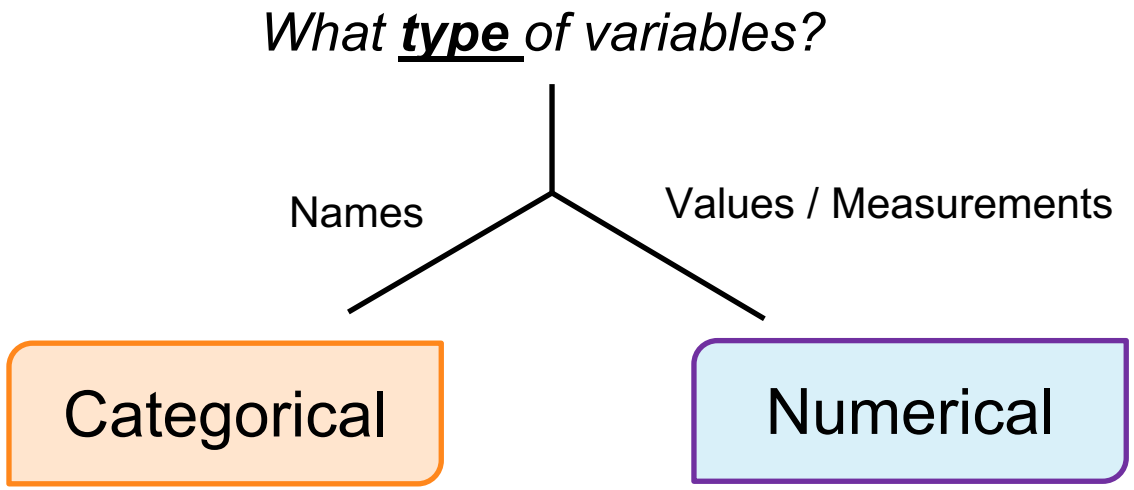
1. Univariate Data

Univariate Data

Types of Data



Types of Variables



- We can break down categorical and numerical variables into multiple categories!

Types of Variables

Categorical Data

Data is divided into *categories*

eg// Eye colour, football team

Nominal Categorical

No sensible way to sort the categories!

Eg// Hair colour:
Does it make sense to order blonde, brown, red and black hair? *No!*

Ordinal Categorical

Categories have a natural order, but the interval between them is not specific.

Eg// How satisfied are you?

- Very Satisfied
- Somewhat Satisfied
- Neutral
- Somewhat Dissatisfied
- Very Dissatisfied

Types of Variables

Numerical Data

Data that you measure or count

Eg// Height, number of students

Numerical Discrete

Data you can *count*, can only take on a finite set of values.

Eg// Number of people at this lecture!
I could count all of you, and I would get a distinct number. Even if I wanted, I couldn't get a more 'accurate' value.

Numerical Continuous

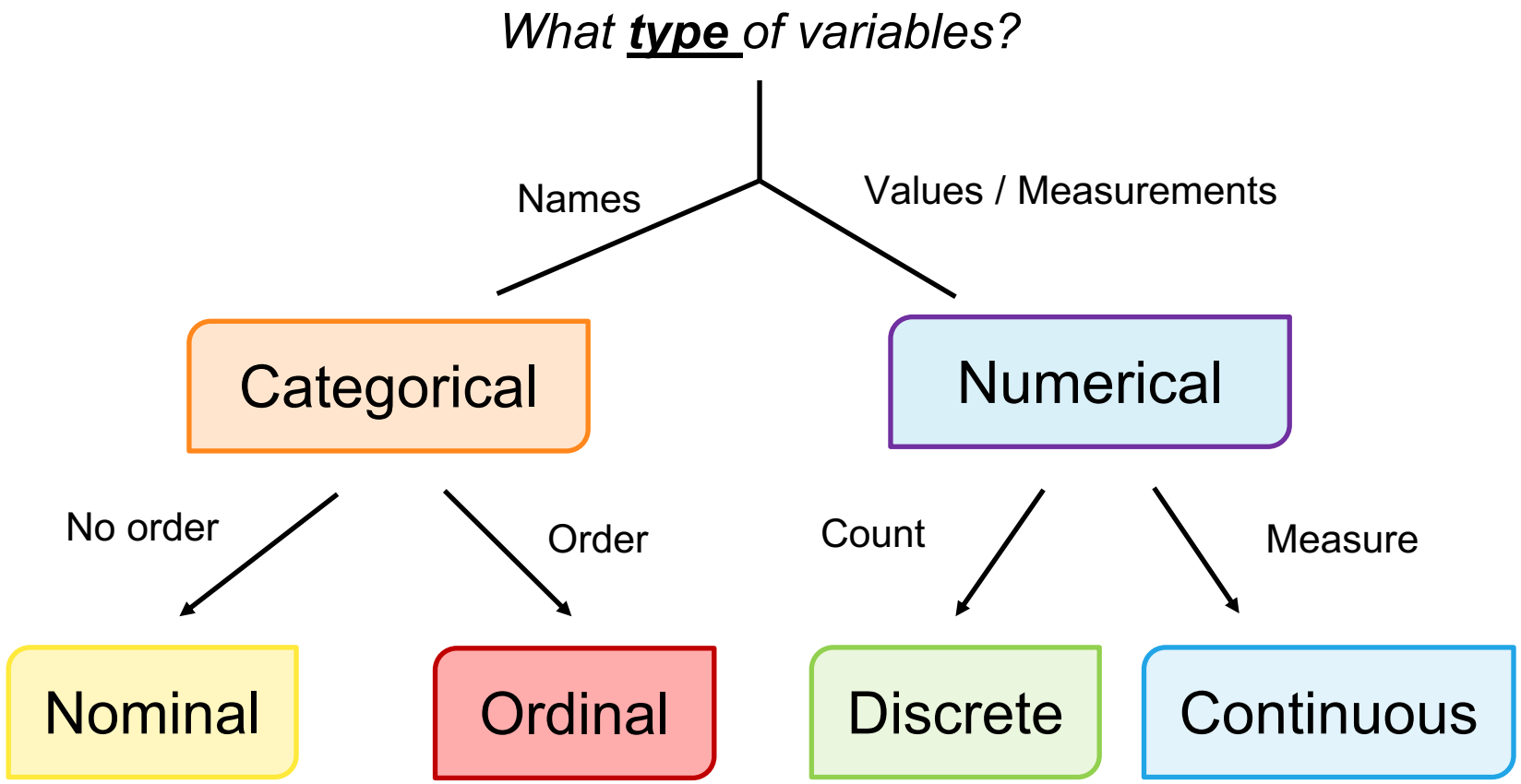
Data that you *measure*, can take any value (infinite possibilities)

Eg// How much does a $2 coin weight?

- 7 grams
- 6.6 grams
- 6.60 grams
- 6.601 grams

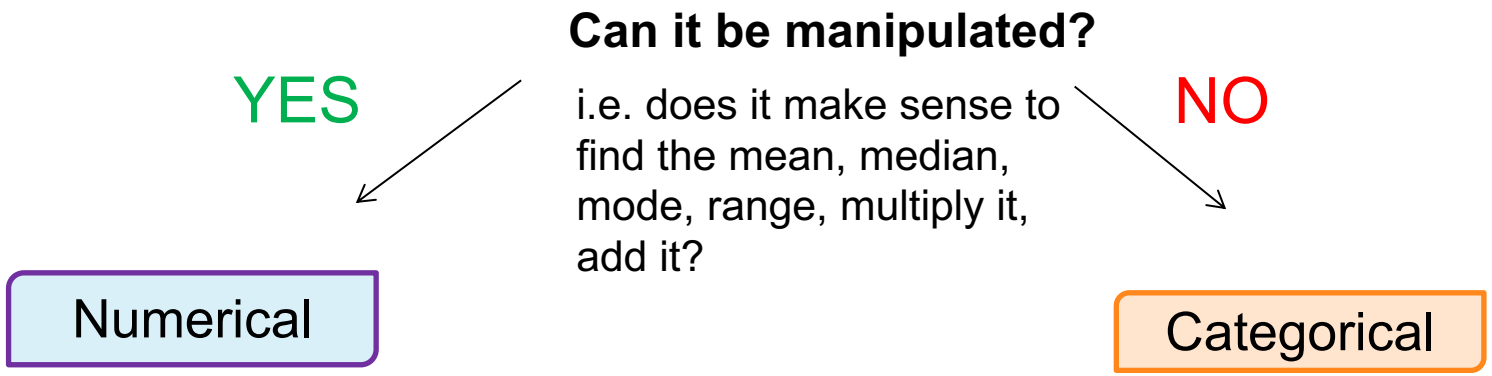
Univariate Data

Types of Variables



Checklist for Variables

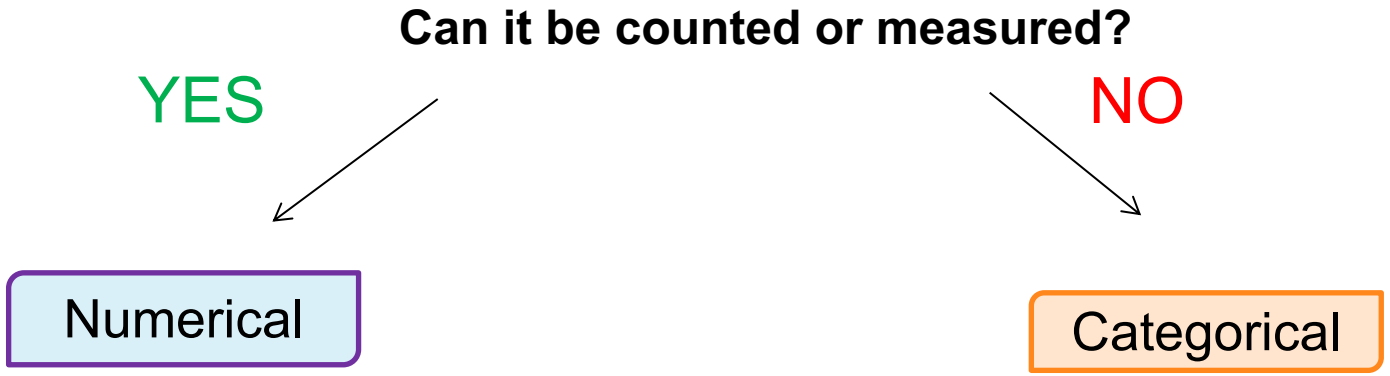
Is the variable categorical or numerical?



- It makes sense to subtract two heights from each other, heights are *numerical*.
- It **doesn't** make sense to subtract two eyes colours from each other, eyes colours are *categorical.*

Checklist for Variables

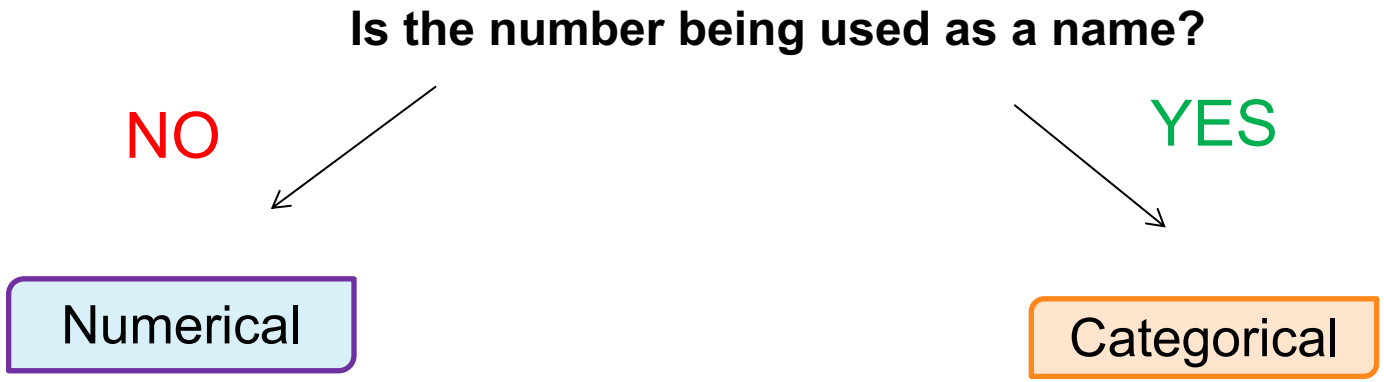
Is the variable categorical or numerical?



Warning: Although numbers usually mean that a variable is numerical, it doesn't always!
Categorical variables can contain numbers too!

Checklist for Variables

Is the variable categorical or numerical?



- For example, numbers on uniforms are used to *identify players,* they act as *names*, and therefore they are categorical.
- Post codes, house numbers and ratings on number scales (e.g. rate out of 5 stars) are other common categorical variables that use numbers!

PRACTICE QUESTIONS

Assign the following variables as numerical, categorical, and whether they are discrete/continuous or ordinal/nominal:

- How often do you study (often, sometimes, rarely)
- The temperature in degrees Celsius
- The cost to fill a car with a tank of petrol
- Shoe size (6, 8, 10)
- Colour of a pencil (red, green, blue)
- Floor levels in a building (1, 2, 3, 4)
- The number of pages in a book

PRACTICE QUESTIONS

Assign the following variables as numerical, categorical, and whether they are discrete/continuous or ordinal/nominal:

- How often do you study (often, sometimes, rarely) — Categorical ordinal
- The temperature in degrees Celsius — Numerical continuous
- The cost to fill a car with a tank of petrol — Numerical discrete
- Shoe size (6, 8, 10) — Categorical ordinal
- Colour of a pencil (red, green, blue) — Categorical nominal
- Floor levels in a building (1, 2, 3, 4) — Categorical ordinal
- The number of pages in a book — Numerical discrete

PRACTICE QUESTION

Blood pressure	Age	
	Under 50 years	50 years or over
low	15	5
normal	32	24
high	11	23
Total	58	52

Question 2

The variables *blood pressure* (low, normal, high) and *age* (under 50 years, 50 years or over) are

A. both nominal variables.

B. both ordinal variables.

C. a nominal variable and an ordinal variable respectively.

D. an ordinal variable and a nominal variable respectively.

E. a continuous variable and an ordinal variable respectively.

VCAA – 2016 Further Math Exam 1 - Data Analysis - Question 2

PRACTICE QUESTION

Blood pressure	Age	
	Under 50 years	50 years or over
low	15	5
normal	32	24
high	11	23
Total	58	52

Question 2

The variables *blood pressure* (low, normal, high) and *age* (under 50 years, 50 years or over) are

A. both nominal variables.

B. both ordinal variables.

C. a nominal variable and an ordinal variable respectively.

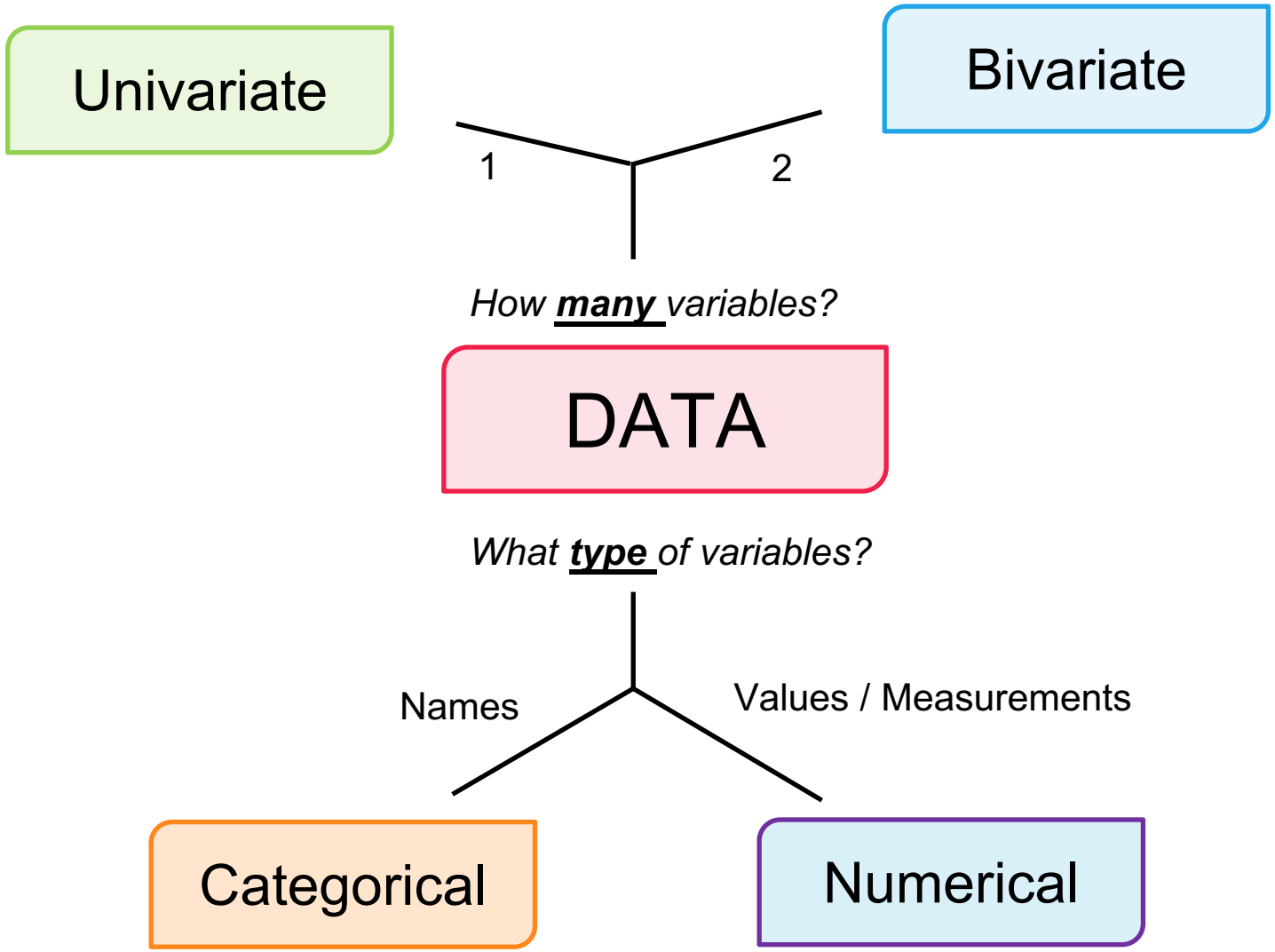
D. an ordinal variable and a nominal variable respectively.

E. a continuous variable and an ordinal variable respectively.

VCAA – 2016 Further Math Exam 1 - Data Analysis - Question 2

Univariate Data

Types of Data



Univariate Data

Univariate

- One variable (uni meaning one)
- Only one thing changes or is manipulated

Colour	Number
Red	3
Black	10
White	8
Silver	5

Bivariate Data

Bivariate

- Two Variables (bi meaning two)
- There are two things that change or are manipulated

	Walk	Bike	Car	Bus
Year 7	5%	12%	71%	12%
Year 8	7%	11%	68%	14%
Year 9	9%	13%	63%	15%
Year 10	9%	17%	59%	15%

- Bivariate data is super interesting, more on this later…

What type of graph do I use? (Univariate Data)

Categorical	Frequency tables
	Percentage frequency tables
	Bar charts

What type of graph do I use? (Univariate Data)

Categorical	Frequency tables
	Percentage frequency tables
	Bar charts
Numerical	Frequency tables
	Dot plots
	Box plots
	Stem and leaf plots
	Histograms

Frequency Tables

Provide the variable and the frequency

Categorical

Eg// Preferred social media platform?

Social media platform	Frequency
Facebook	26
Twitter	19
Instagram	20
Snapchat	17
TikTok	1
Total	**83**

Percentage Frequency Tables

Frequency may also be displayed as percentage frequency

Categorical

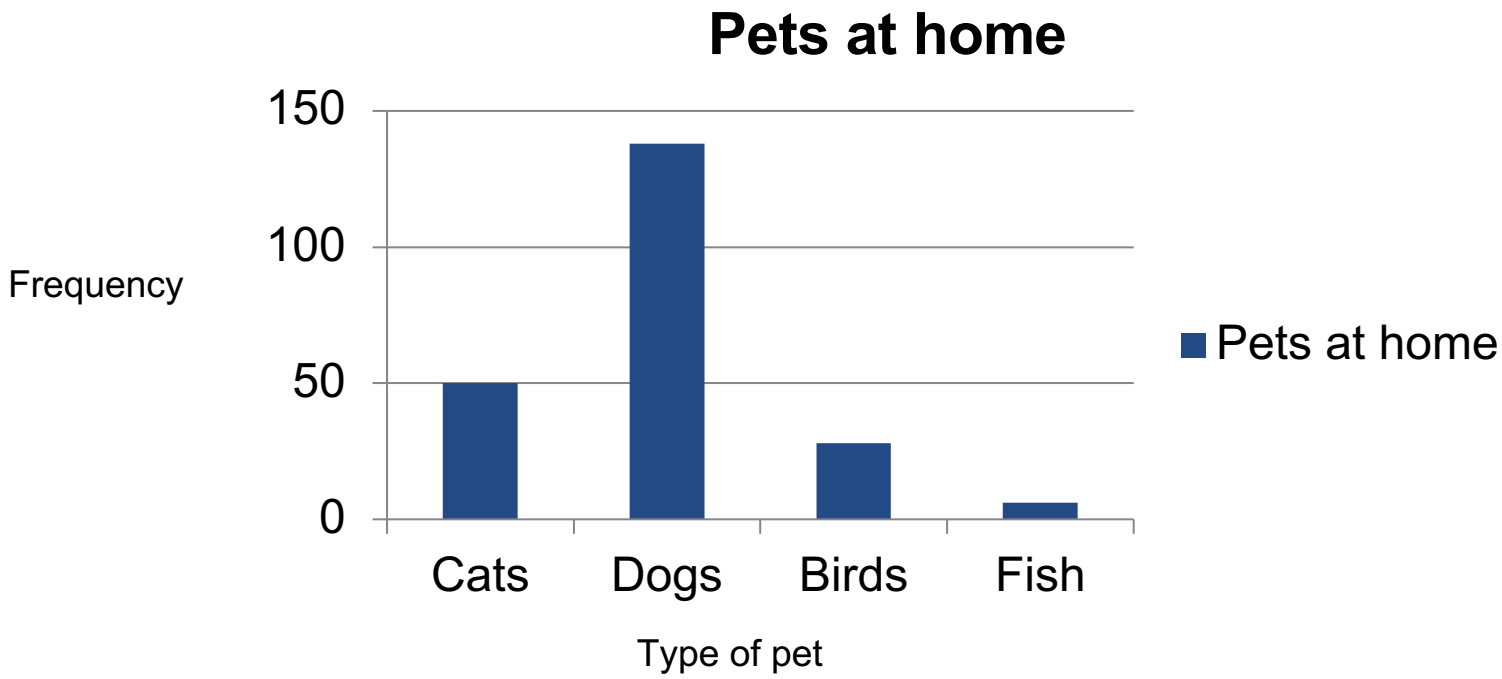
Eg// Preferred social media platform?

Social media platform	% Frequency
Facebook	31.33%
Twitter	22.89%
Instagram	24.10%
Snapchat	20.48%
TikTok	1.20%
Total	**100%**

Bar Charts

- Variable on the x-axis, frequency on the y-axis
- Label axes, must rule lines
- Bars of equal width with space between them

Categorical



Describing Categorical Data

- Explain/Describe paragraphs pop up every now and then
- Have a template in your summaries
- When answering these kinds of questions, you *must*:
  - Summarise the context
  - Identify the mode (also called modal category, dominant category)
  - Quote its frequency
  - Quote other frequencies of interest

Example Question

Q: Comment on the data shown in the frequency table below.

Climate	Frequency	% Frequency
Hot	6	26.1%
Mild	14	60.9%
Cold	3	13.0%
Total	**23**	**100**

The climate types of 23 countries were classified as being “cold”, “mild” or “hot”. The majority of the countries, 60.9%, were found to have a mild climate. Of the remaining countries, 26.1% were found to have a hot climate, while 13% were found to have a cold climate.

Univariate Data

Example Question

Climate	Frequency	% Frequency
Hot	6	26.1%
Mild	14	60.9%
Cold	3	13.0%
Total	**23**	**100**

Context

The climate types of 23 countries were classified as being “cold”, “mild” or “hot”. The majority of the countries, 60.9%, were found to have a mild climate. Of the remaining countries, 26.1% were found to have a hot climate, while 13% were found to have a cold climate.

Example Question

Climate	Frequency	% Frequency
Hot	6	26.1%
Mild	14	60.9%
Cold	3	13.0%
Total	**23**	**100**

Mode

The climate types of 23 countries were classified as being “cold”, “mild” or “hot”. The majority of the countries, 60.9%, were found to have a mild climate. Of the remaining countries, 26.1% were found to have a hot climate, while 13% were found to have a cold climate.

Example Question

Climate	Frequency	% Frequency
Hot	6	26.1%
Mild	14	60.9%
Cold	3	13.0%
Total	**23**	**100**

Note other frequencies

The climate types of 23 countries were classified as being “cold”, “mild” or “hot”. The majority of the countries, 60.9%, were found to have a mild climate. Of the remaining countries, 26.1% were found to have a hot climate, while 13% were found to have a cold climate.

FREQUENCY TABLES

- Variable and frequency displayed
- Frequency can be real or percentage frequency
- Data with a large spread may be displayed as “grouped” data

Numerical

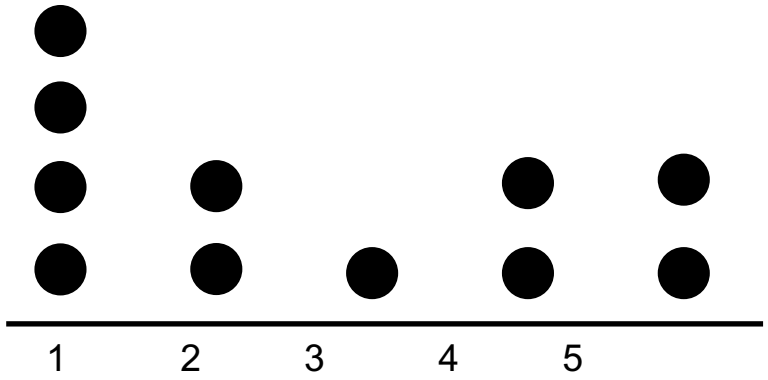
Age	Frequency
17	10
18	13
19	2

Age	Frequency
0-20	18
20-40	43
40-60	25

DOT PLOTS

Numerical

- Used for *discrete* numerical data
- Only the x-axis is used, displaying the variable
- Number of dots represents frequency

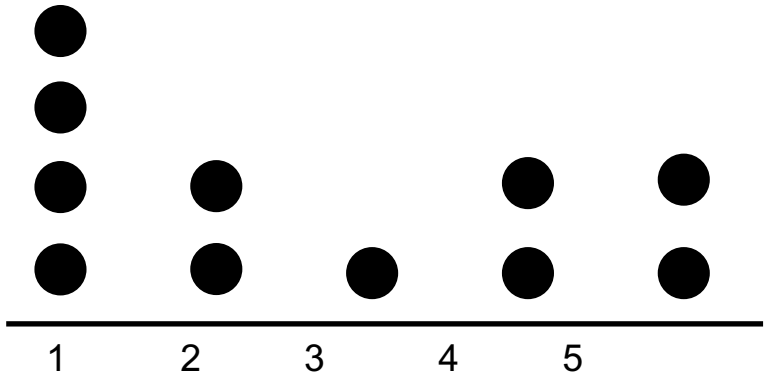


DOT PLOTS

Numerical

Common question:

- Finding the median
  - Count the total
  - Find the dot denoting the middle point of the data

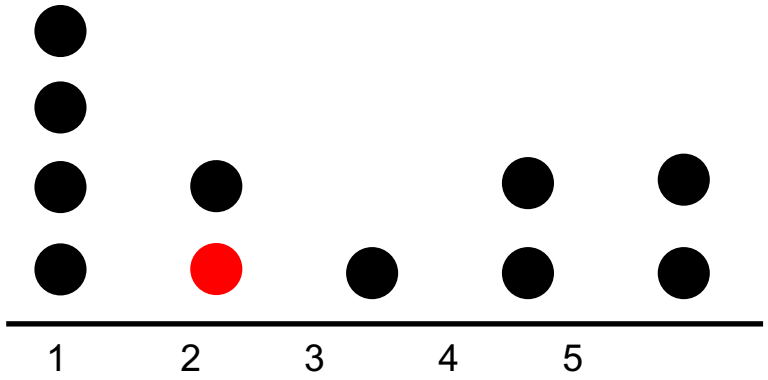


DOT PLOTS

Numerical

Common question:

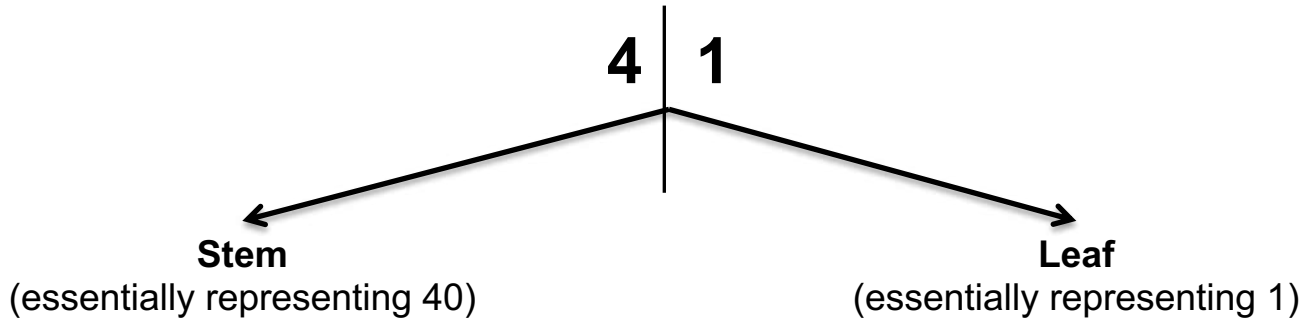
- Finding the median
  - Count the total
  - Find the dot denoting the middle point of the data



STEM AND LEAF PLOTS

Numerical

- There are **two basic parts:**
  - Stem: the first digit(s)
  - Leaf: the last digit(s)
- For example, the number 41 may be shown as follows:



Stem and Leaf Plots

Numerical

Warning: **Always** include a key/legend with your stem plot.

Key:

1 | 0 = 10

Stem	Leaf
1	0 2
2	2 3 8 9
3	9 9
4	4 4 7
5	
6	0 1 9

10

Stem and Leaf Plots

Numerical

Warning: **Always** include a key/legend with your stem plot.

Key:

1 | 0 = 100

Stem	Leaf
1	0 2
2	2 3 8 9
3	9 9
4	4 4 7
5	
6	0 1 9

100

Stem and Leaf Ploys

Numerical

Warning: **Always** include a key/legend with your stem plot.

- The leaf is ordered
- Can split the stem up in half or fifths if plot is bunched

Key:

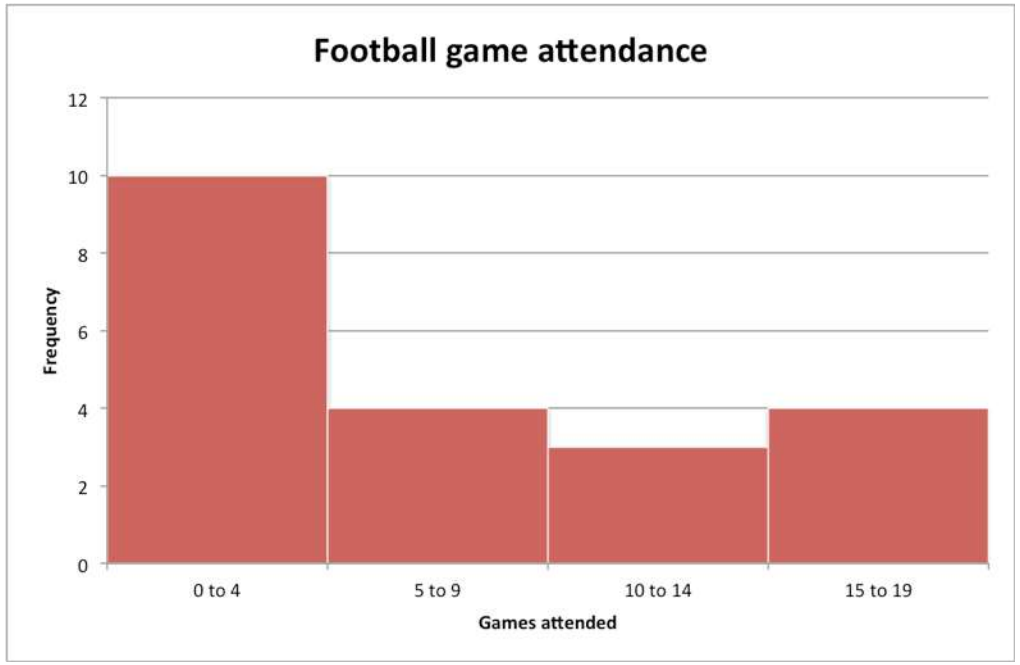
1 | 0 = 10

Stem	Leaf
1	0 2
2	2 3 8 9
3	9 9
4	4 4 7
5	
6	0 1 9

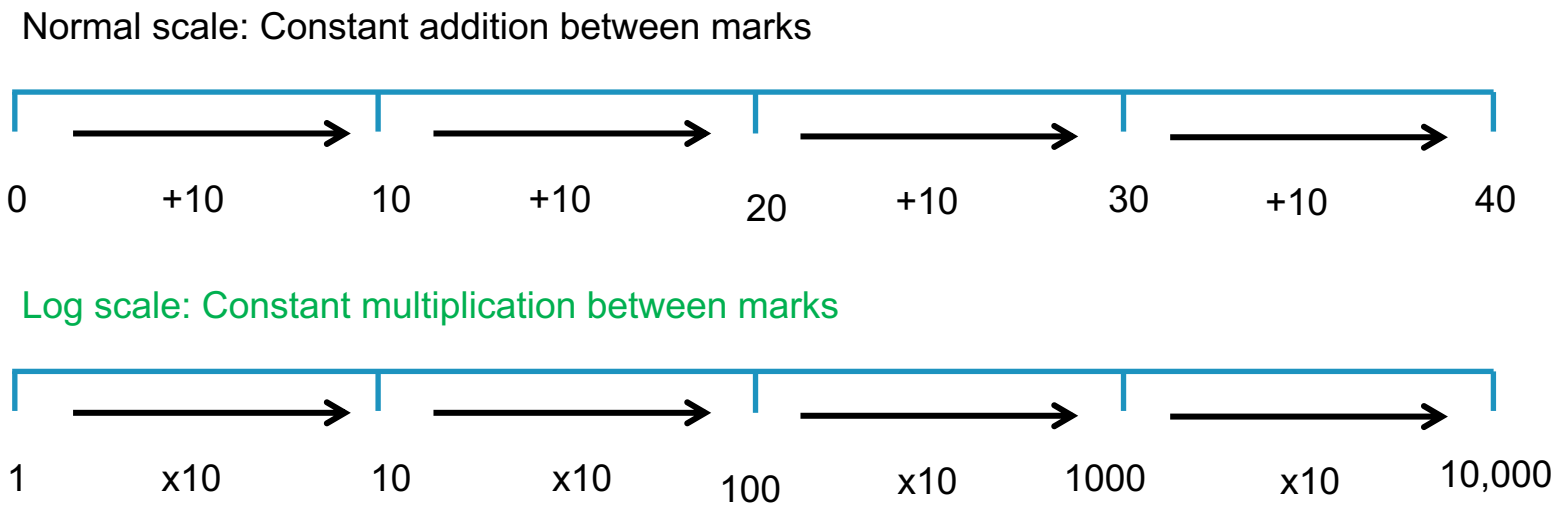
Histograms

Numerical

- Similar to bar charts, but *no space between columns*
- Variable on x-axis, frequency on y-axis
- Useful to identify key features of data that we use in descriptions (shape, spread, centre, outliers)



- Sometimes, the data's range is too large to display on a regular histogram
- In these cases, we use log scale histograms as a solution!
- wth is a log scale?



Properties of logs:

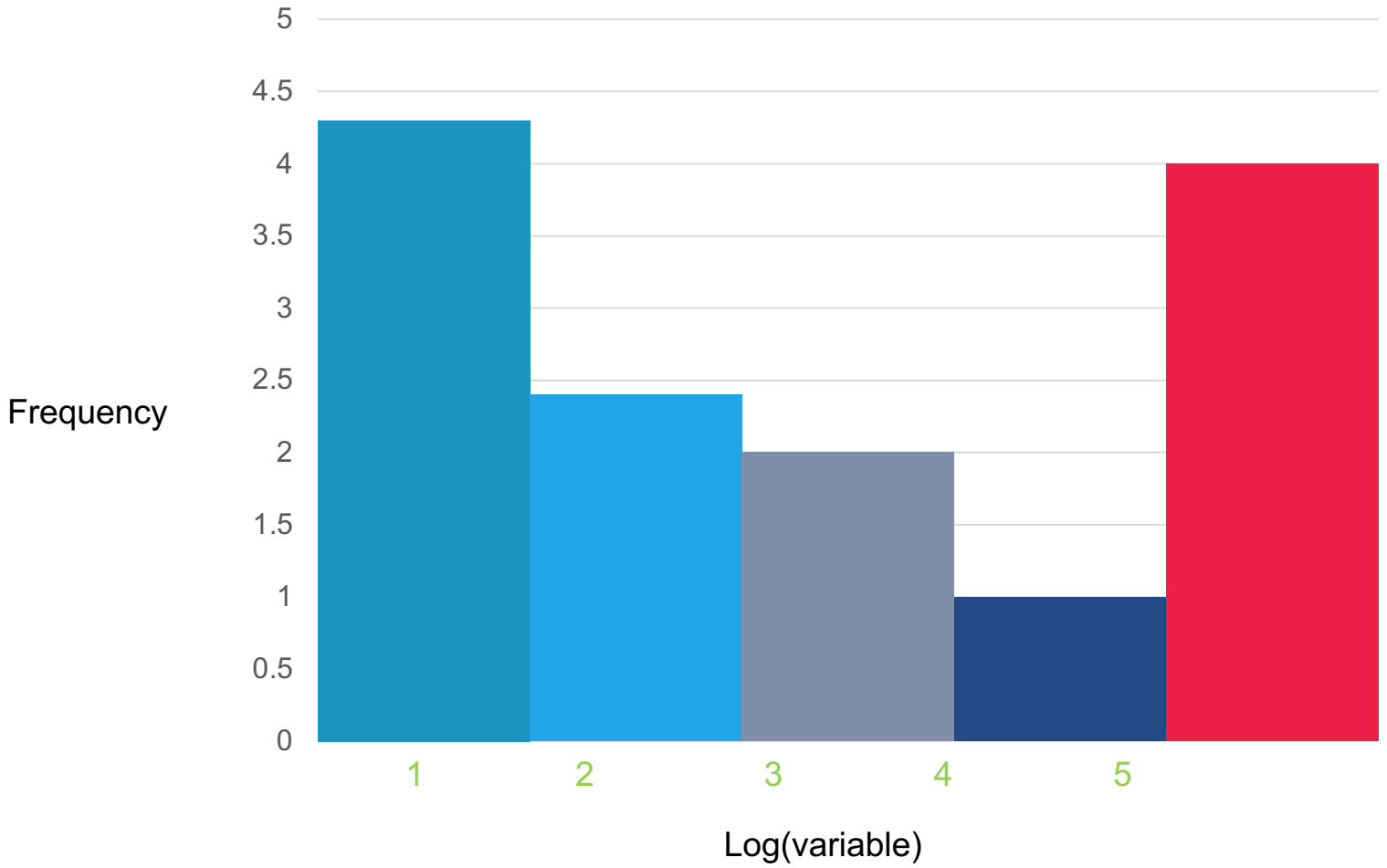
- If a number is greater than 1 its log is greater than 0
- If a number is greater than 0 but less than 1 its log is negative
- If a number is 0 its log is undefined, and you can't have logs of negative numbers!

Warning: When displaying logs on an axis we only use their order of magnitude (10^2 becomes 2), though we must label the axis as *log(variable)*.

Univariate Data

Log Scale Histogram

Eg//



Log Scale Guide

A handy log guide:

0.01	0.1	1	10	100	1000	10000	100000	1000000
10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}

Converting Logs

- To find the log of a number
  - Eg. What is the log of 150?
  - $\log_{10}(150) = 2.176$

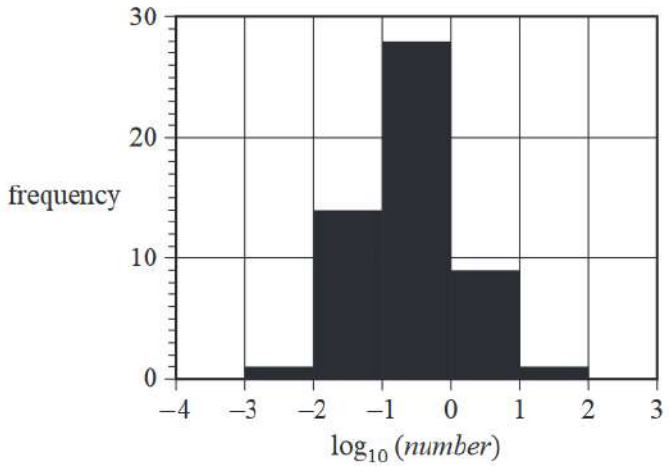
Use calculator!!

- To find the number of a log
  - Eg. Find the number of log 1.683
  - $10^{1.683} = 48.1948$

PRACTICE QUESTION

Question 7

The histogram below shows the distribution of the *number* of billionaires per million people for the same 53 countries as in Question 6, but this time plotted on a $\log_{10}$ scale.



Data: Gapminder

Based on this histogram, the number of countries with one or more billionaires per million people is

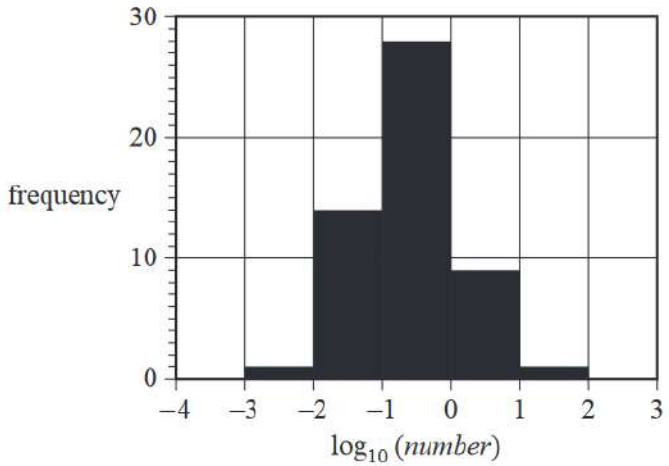
A. 1
B. 3
C. 8
D. 9
E. 10

VCAA 2016 Exam 1 – Question 7

PRACTICE QUESTION

Question 7

The histogram below shows the distribution of the *number* of billionaires per million people for the same 53 countries as in Question 6, but this time plotted on a $\log_{10}$ scale.



Data: Gapminder

Based on this histogram, the number of countries with one or more billionaires per million people is

A. 1
B. 3
C. 8
D. 9
E. 10

VCAA 2016 Exam 1 – Question 7

5 Figure Summary

Includes:

- The minimum value
- The value of quartile 1 (Q1)
- The median
- The value of quartile (Q3)
- The maximum value

- We can work this out by hand or on the calculator, depending on what set of data you have either one may be quicker!

5 Figure Summary – Example

How can we find the 5 number summary by hand?

2 5 7 8 9 13 14 15 16 20 21 25 37 41

5 Figure Summary – Example

How can we find the 5 number summary by hand?

2 5 7 8 9 13 14 15 16 20 21 25 37 41

We can see that we have **14** numbers here

5 Figure Summary – Example

How can we find the 5 number summary by hand?

2 5 7 8 9 13 14 | 15 16 20 21 25 37 41

We can see that we have **14** numbers here

Median: 14.5

5 Figure Summary – Example

How can we find the 5 number summary by hand?

2 5 7 8 9 13 14 | 15 16 20 21 25 37 41

Q1: 8
Median: 14.5

5 Figure Summary – Example

How can we find the 5 number summary by hand?

2 5 7 8 9 13 14 | 15 16 20 21 25 37 41

Q1: 8
Median: 14.5
Q3: 21

5 Figure Summary – Example

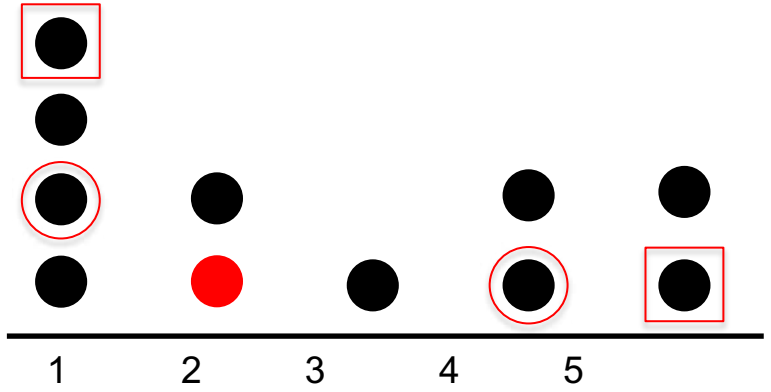
How can we find the 5 number summary by hand?

2 5 7 8 9 13 14 | 15 16 20 21 25 37 41

Minimum: 2
Q1: 8
Median: 14.5
Q3: 21
Max: 41

5 Figure Summary – Dot Plots

How can we find the 5 number summary from a dot plot?



5 Figure Summary – Stem and Leaf

How can we find the 5 number summary from a stem plot?

Stem	Leaf
1	0 2
2	2 3 8 9
3	9 9
4	4 4 7
5	
6	0 1 9

Interquartile Range

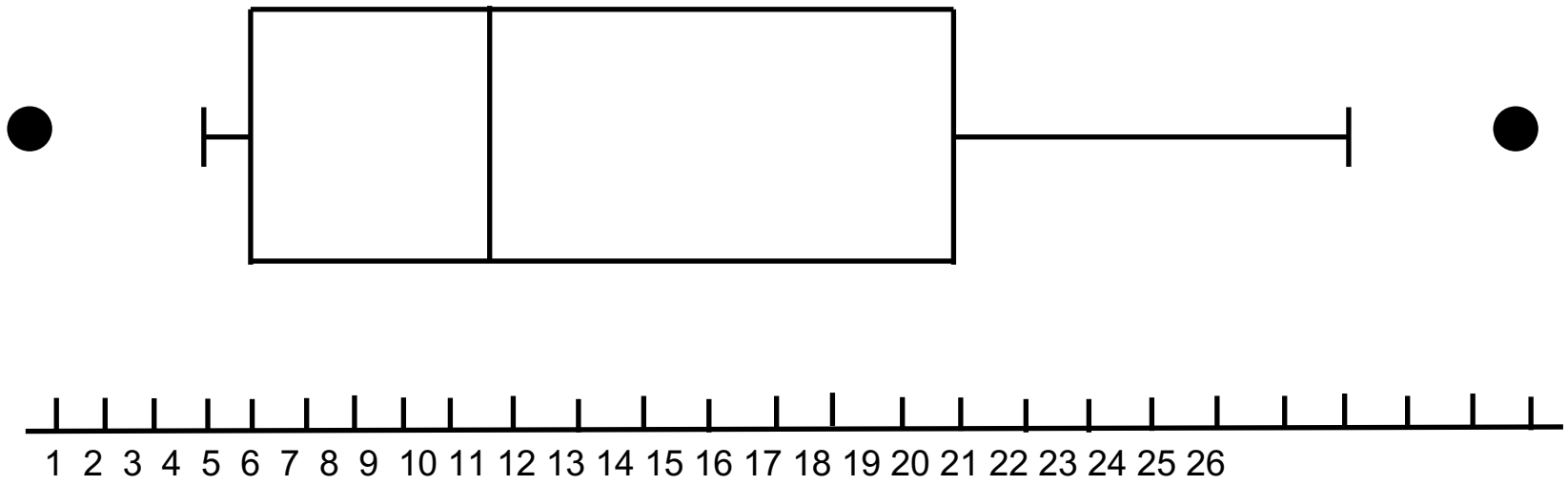
What can we do with these details?

- IQR
  - IQR = Q3 – Q1
- Outlier calculations - Extremely common exam question!
  - Lower fence value
    - Q1 – 1.5 x IQR
  - Upper fence value
    - Q3 + 1.5 x IQR

5 Figure Summary – Box Plots

Visual display of the 5 number summary

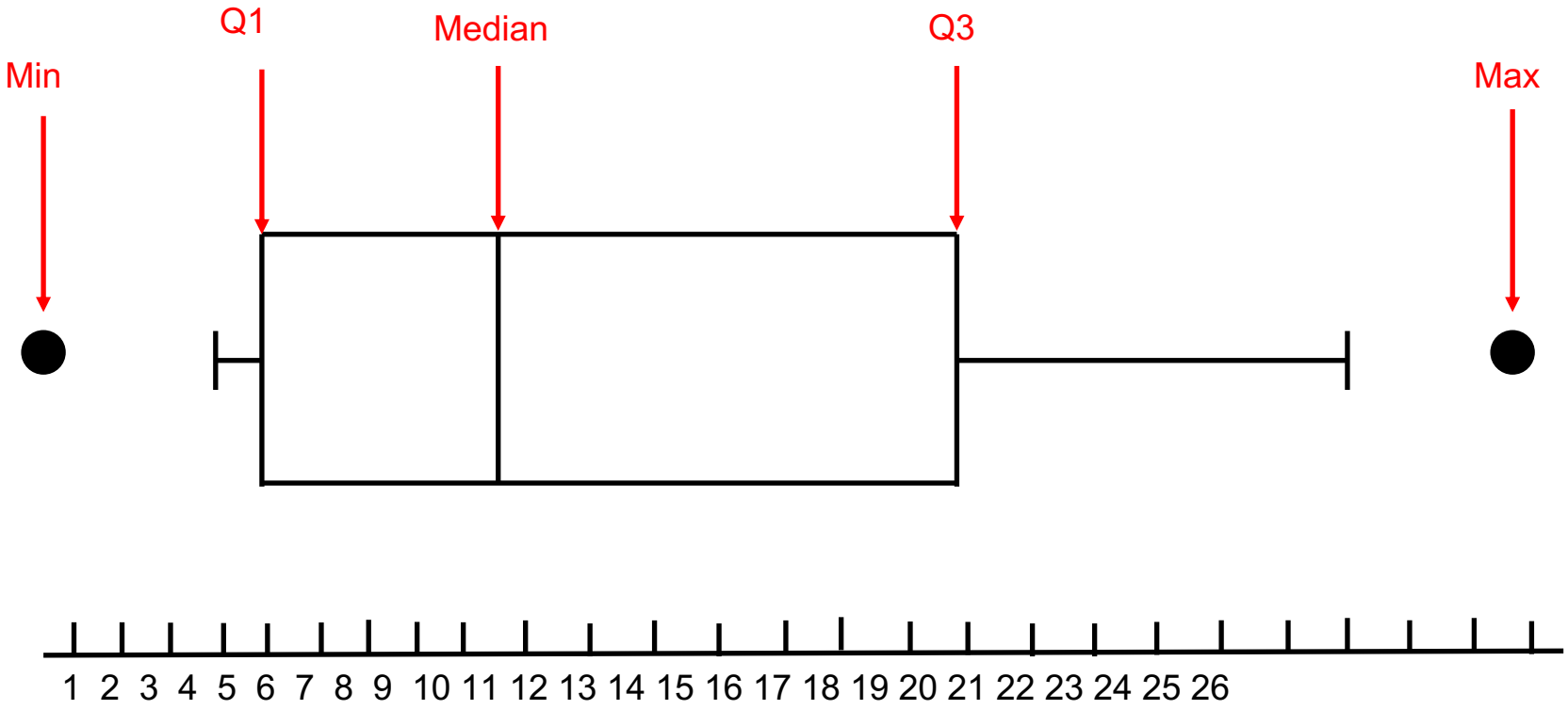
Numerical



5 Figure Summary –Box Plots

Visual display of the 5 number summary

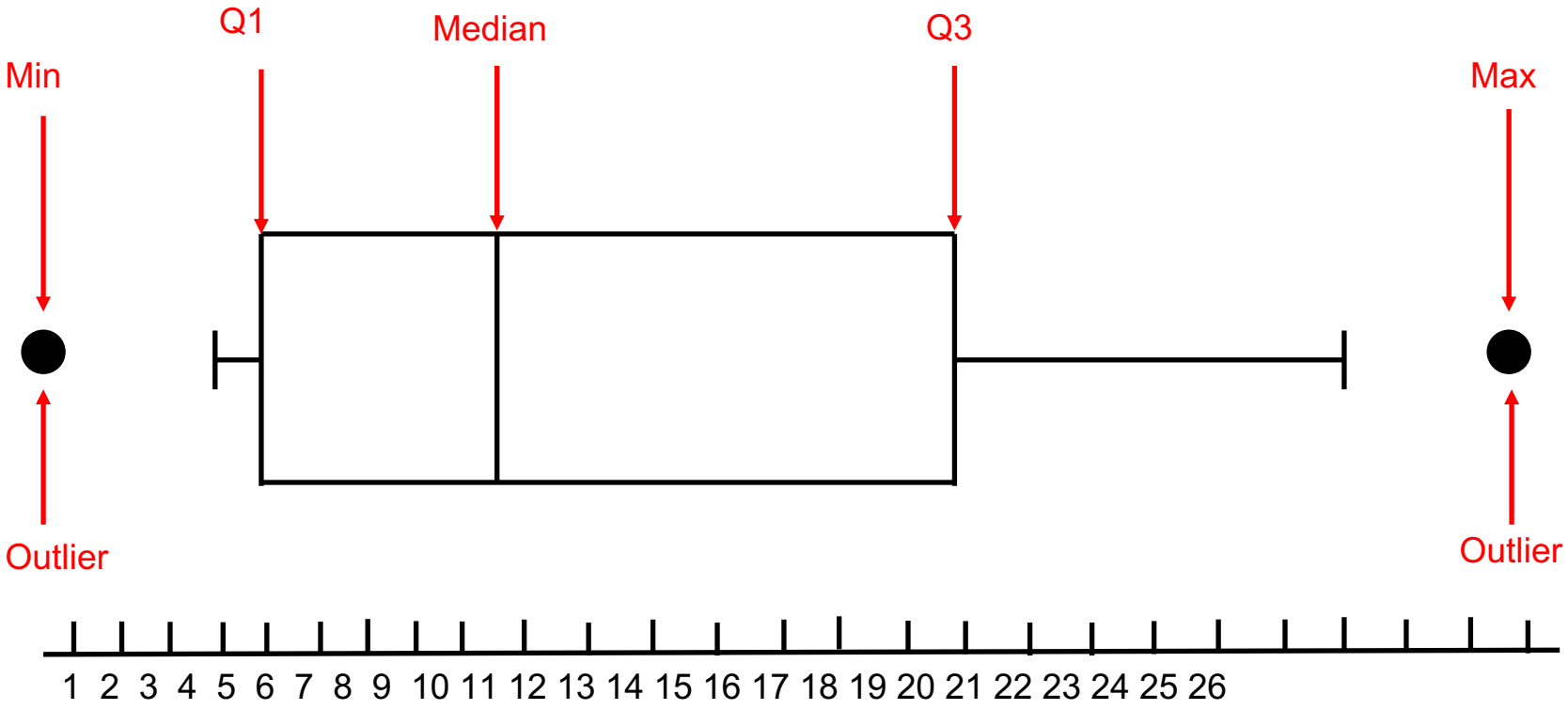
Numerical



5 Figure Summary – Box Plots

Visual display of the 5 number summary

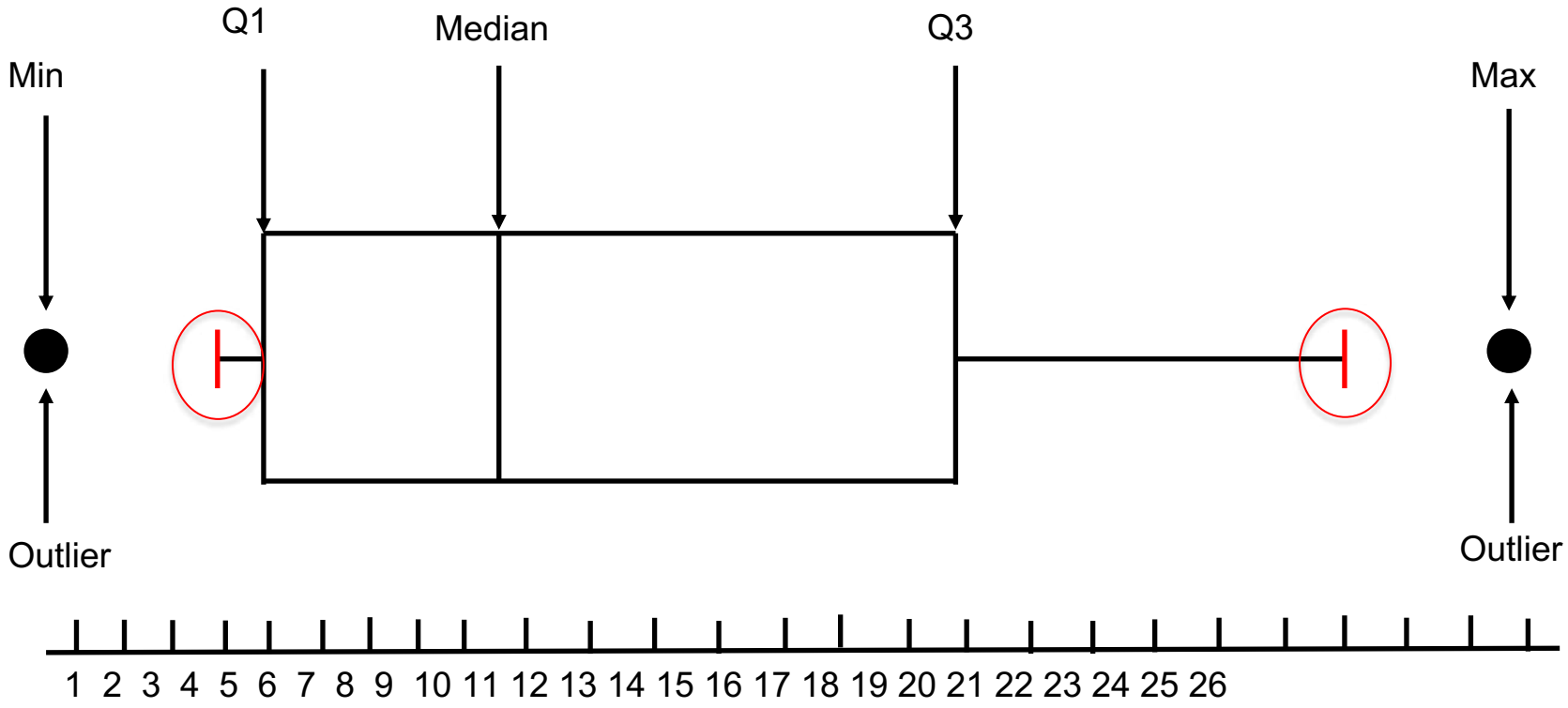
Numerical



5 Figure Summary – Box Plots

Visual display of the 5 number summary

Numerical



Describing Numerical Data

Must discuss: shape, centre, spread and outliers

Outliers: Are there any present? If so, what are they? Also note if there are no outliers.

Centre: Note the mean or the median, perhaps both

- Mean, median, mode

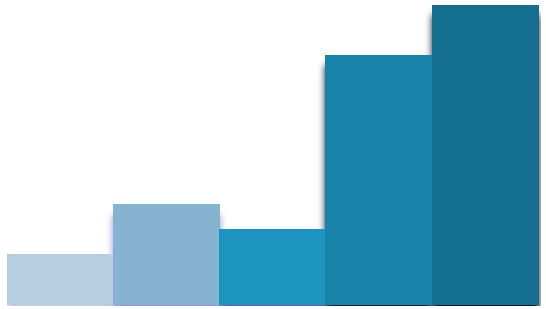
Spread: What is the range of the data

- IQR, range, standard deviation

Shape of Data

Shape:

Positively Skewed



Negatively Skewed

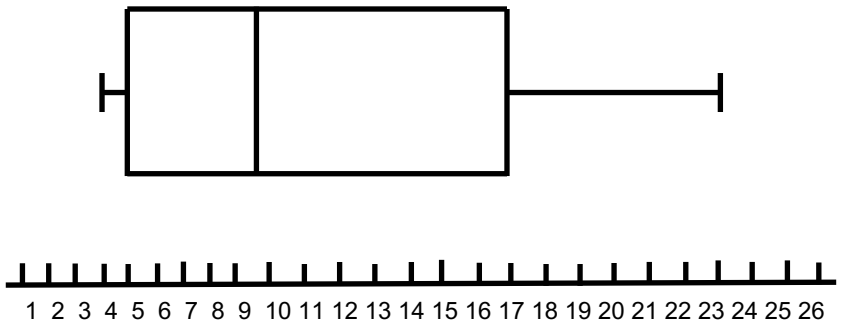
Approximately symmetrical

Bimodal

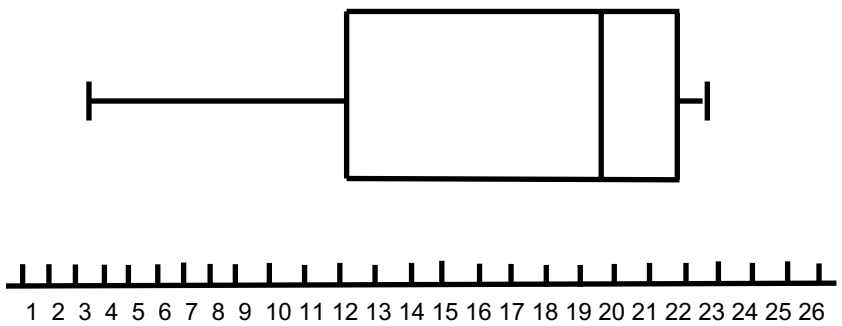
Box Plot Shape

Once again, we look at outliers, centre, spread and shape

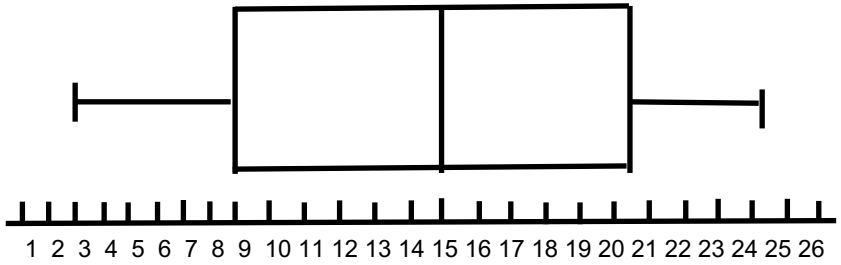
For box plots, shape is displayed as:



Positively Skewed



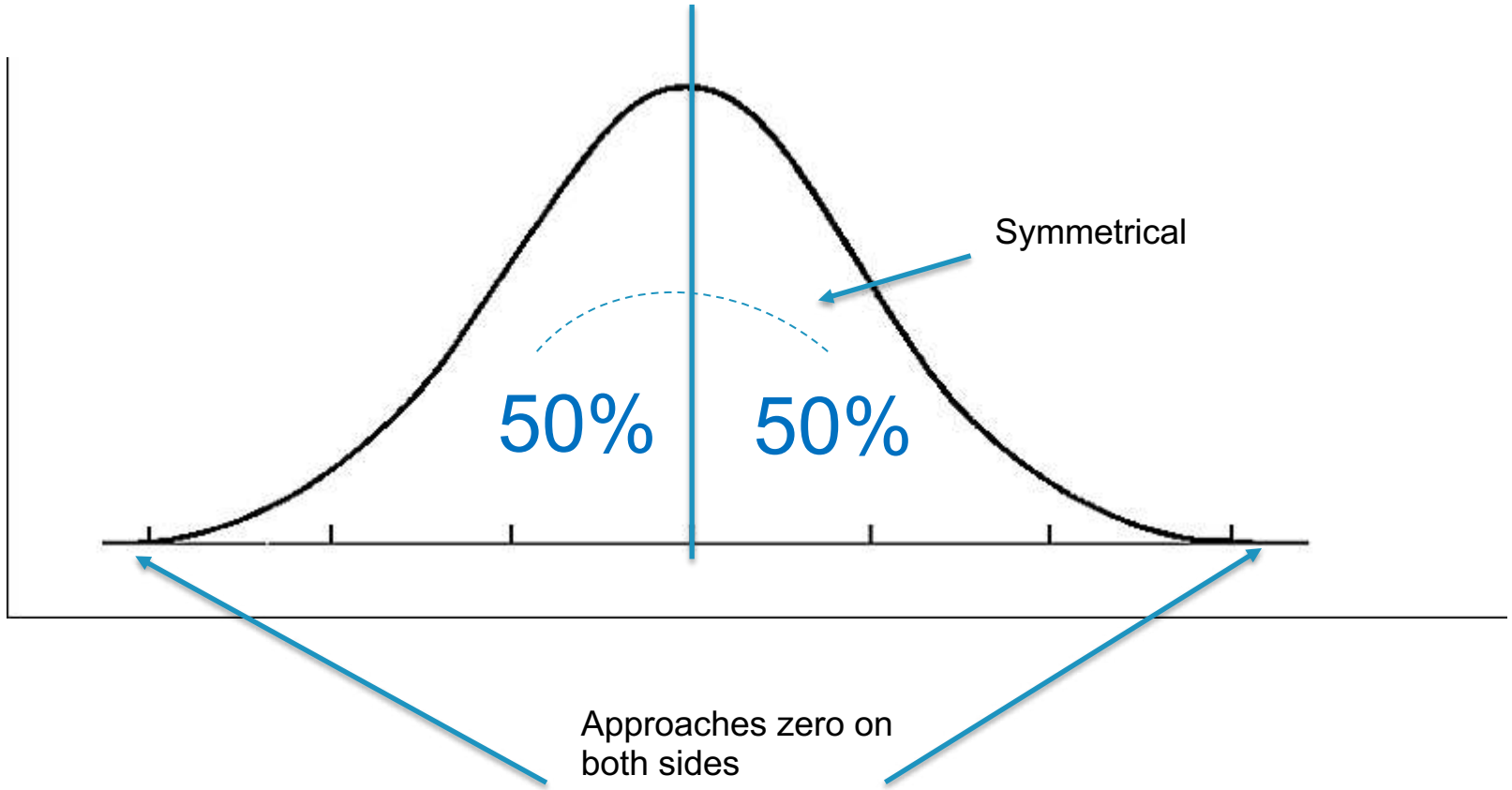
Negatively Skewed



Approximately Symmetrical

Normal Distribution

$$mean = median = mode$$



Standard Deviation

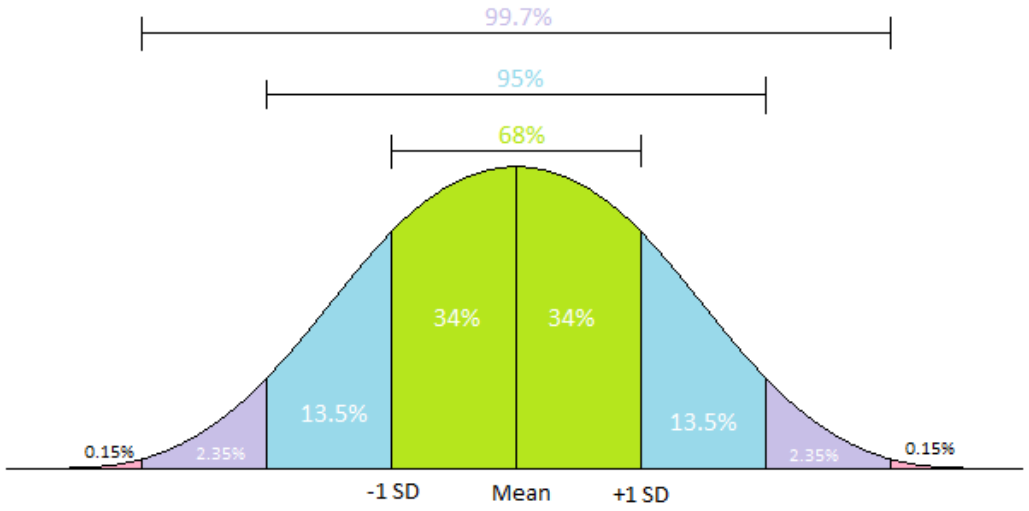
Describes the *spread* of data values around the mean

e.g. mean = 5 $s_x = 2$

- 1 S.D. above mean = 7
- 2 S.D. above mean = 9
- 1 S.D. *below* mean = 3

Standard Deviation

- Around 68% of the data values lie within one standard deviation of the mean.
- Around 95% of the data values lie within two standard deviations of the mean.
- Around 99.7% of the data values lie within three standard deviations of the mean.



PRACTICE QUESTION

Q1. A class of 24 students receives their science test results, with a mean of 32 and a standard deviation of 2.
How many students received a score between 28 and 32?

Q1. A class of 24 students receives their science test results, with a mean of 32 and a standard deviation of 2.
How many students received a score between 28 and 32?

32– 28 = 4
4/2 = 2
2 std deviations under the mean
34% + 13.5% = 47.5%

24 x 0.475 = 11.4
11 students

$$z = \frac{x - \bar{x}}{s}$$

z = z-score

x = actual score

$\bar{x}$ = mean

s = standard deviation

PRACTICE QUESTION

Q1. A class of 24 students receives their science test results, with a mean of 32 and a standard deviation of 2.
How many students received a score between 28 and 34?

Q2. Ben achieved a result of 35, what is his standardised score?

Q1. A class of 24 students receives their science test results, with a mean of 32 and a standard deviation of 2.
How many students received a score between 28 and 34?

Q2. Ben achieved a result of 35, what is his standardised score?

Answers:

Q1. 81.5%

Q2. $\frac{35 - 32}{2} = 1.5$

Summary

Types of data

Categorical	Frequency tables
	Percentage frequency tables
	Bar charts
Numerical	Frequency tables
	Dot/Box plots
	Stem and leaf plots
	Histograms

Displaying data

5 Figure Summary
Shape, centres, spread
68-95-99.7% rule
z - scores

Analysing data

ATARNotes

2. Bivariate Data

Why Bivariate Data?

- Univariate data is great at telling us the ***what***
  - What is the average height of people in this room?
  - What is the most popular colour?
  - What is the average temperature in Melbourne?

- ***Bivariate data*** allows us to ***compare*** data, and focus on the ***why***
  - What is the relationship between age and height?
  - Does gender play a role in someone's favourite colour?
  - How do the average temperatures in all major Australian cities compare?

Explanatory vs Response Variables

- When we've got more than one variable, we give the variables different names.
- Science kids, you'll know these as independent and dependent variables.
- Here, we call them ***explanatory*** and ***response*** variables.

Explanatory vs Response Variables

Explanatory variable → **Response variable**

- Also known as IV or EV
- The variable that you expect when changed, will ‘explain’ to some extent the change in another variable.
- Plotted on x axis

- Also known as DV or RV
- The variable that you think will be changed ‘as a response’ to a changing EV.
- Plotted on y axis.

Age ⇄ Shoe size

Displaying Bivariate Data

Type of Data		Graph
Explanatory Variable	Response Variable	
Categorical	Categorical	Segmented Bar Chart

Displaying Bivariate Data

Type of Data		Graph
Explanatory Variable	Response Variable	
Categorical	Categorical	Segmented Bar Chart
Numerical	Categorical	Parallel Box Plots

Displaying Bivariate Data

Type of Data		Graph
Explanatory Variable	Response Variable	
Categorical	Categorical	Segmented Bar Chart
Numerical	Categorical	Parallel Box Plots
Numerical	Categorical (2 categories)	Back to Back Stem and Leaf Plots

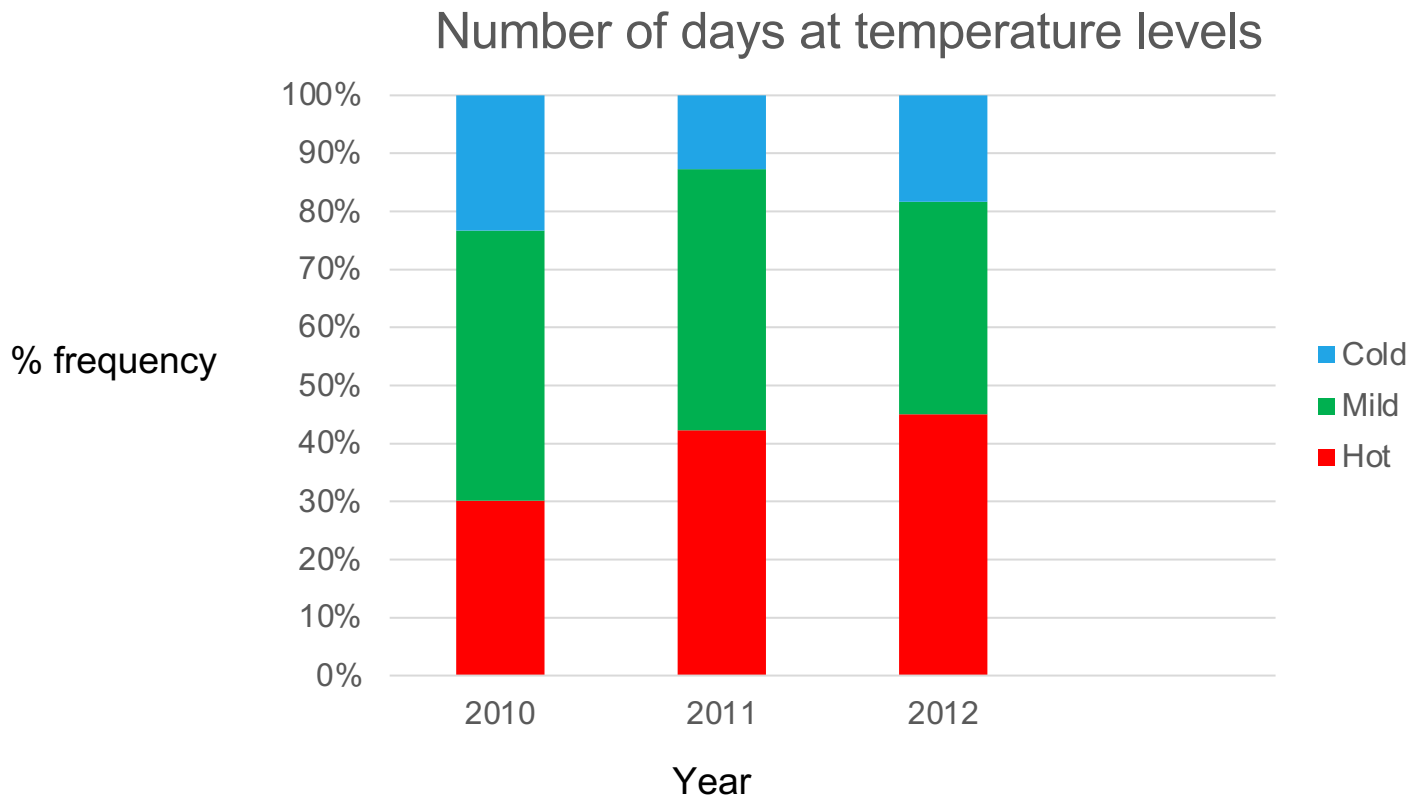
Displaying Bivariate Data

Type of Data		Graph
Explanatory Variable	Response Variable	
Categorical	Categorical	Segmented Bar Chart or Two Way Frequency Table
Numerical	Categorical	Parallel Box Plots
Numerical	Categorical (2 categories)	Back to Back Stem and Leaf Plots
Numerical	Numerical	Scatterplots

Segmented Bar Charts

Categorical | Categorical

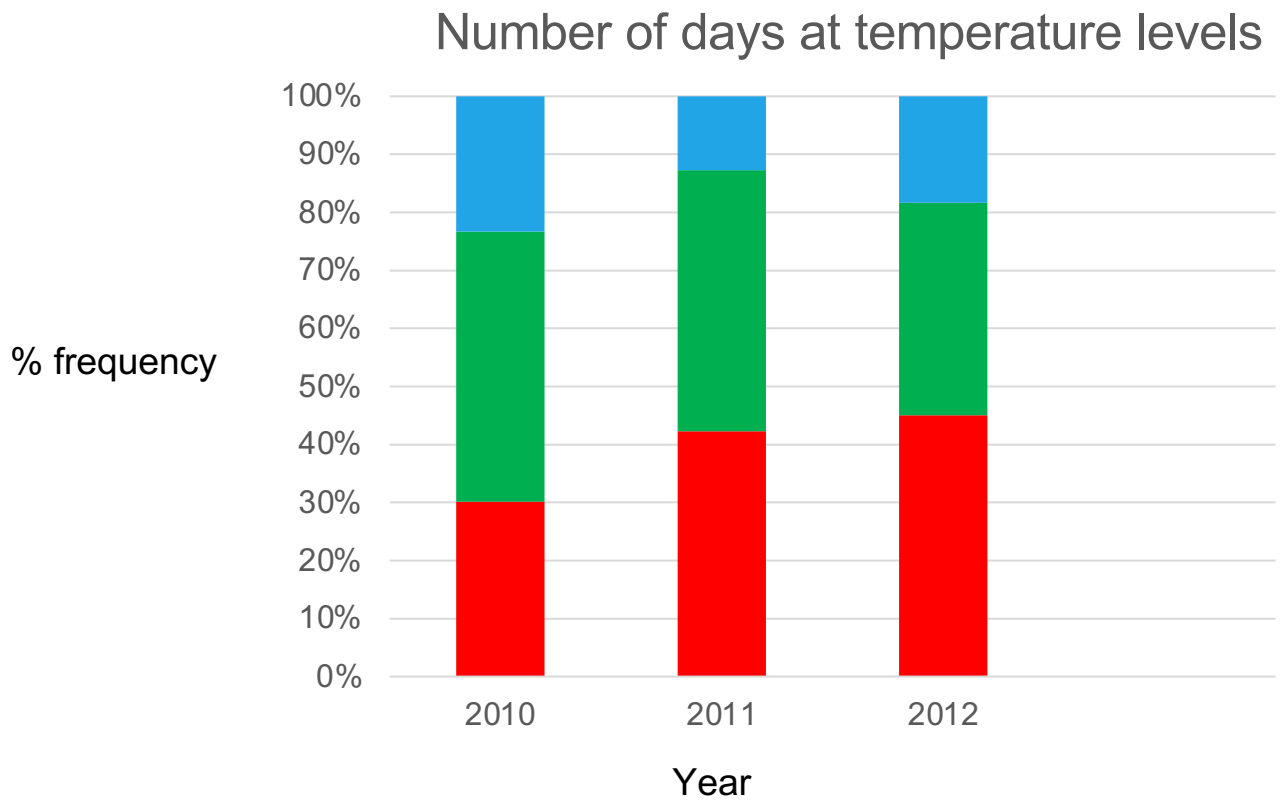
- Variable on x-axis, frequency on y-axis
- Can be in terms of raw number or percentage



SEGMENTED BAR CHARTS

Categorical | Categorical

Note: Ensure to include key!



Two Way Frequency Tables

Categorical | Categorical

Two-Way Frequency Table

- RV = Rows
- EV = Columns

Attitude	Year Level	
	11	12
For	36%	81%
Against	64%	19%
Total	100%	100%

Two Way Frequency Tables

Categorical | Categorical

- What can we see from this?
- There seems to be an association
- If it was random, we would expect percentages to be around 50/50, but they're not!

Attitude	Year Level	
	11	12
For	36%	81%
Against	64%	19%
Total	100%	100%

Back to Back Stem and Leaf Plots

Numerical | Categorical

Back-to-back stem and leaf plots

Blue eyes		Brown eyes
8	5	
1	6	0 2 4 6 9
8 3 0	7	1 3 5
5 4 2 0	8	5
9 6 2	9	9
	10	0 0 0

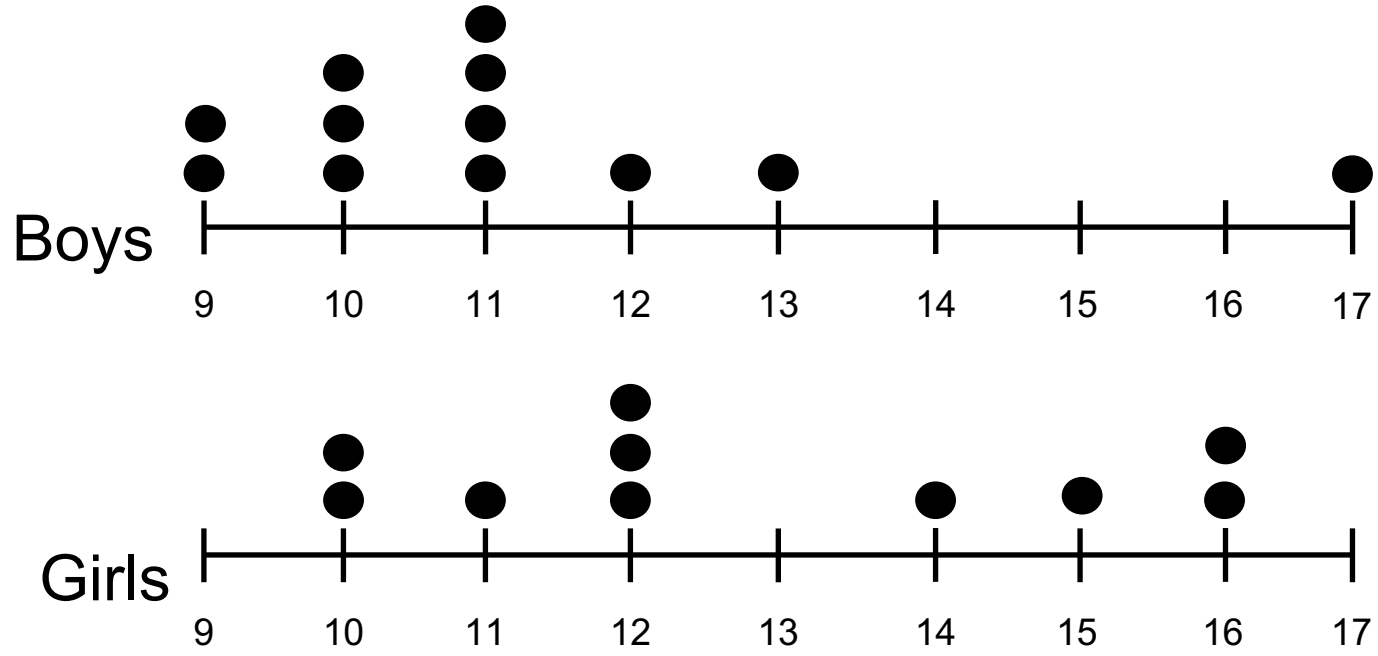
Key

1 | 0 = 10

Parallel Dot Plots

Numerical | Categorical

- One categorical variable, one numerical variables
- Allows for easy comparison of distributions' shape



Comparing Distribution

Can be asked to compare data sets looking at graphs

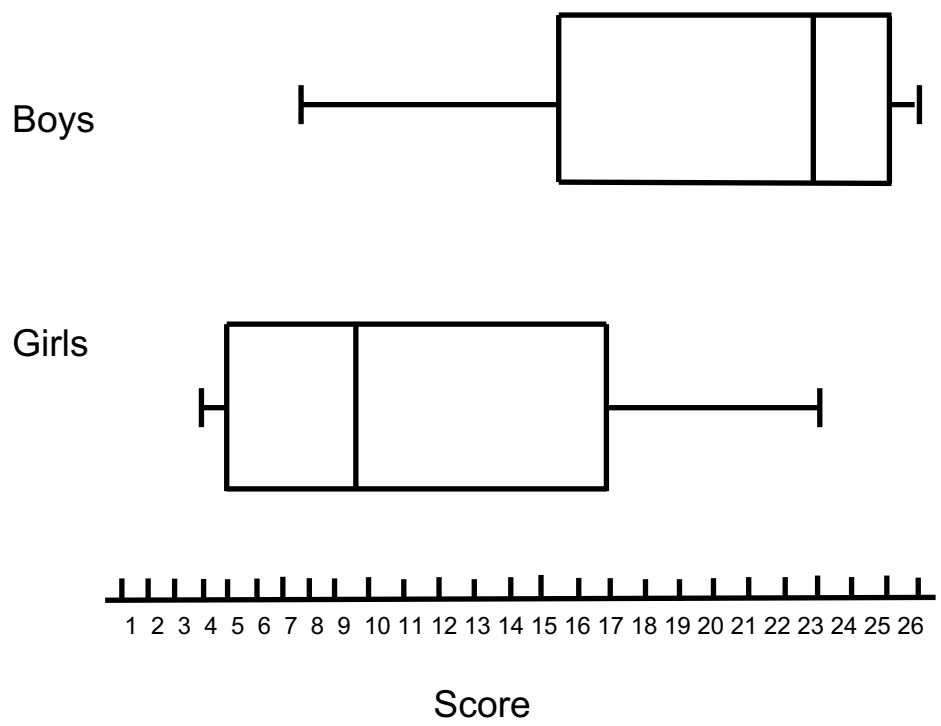
- Box Plots
- Histograms
- Dot Plots
- Back to Back Stem and Leaf Plots

We look at:

- Centre
- Spread
- Shape

Comparing Distribution – Example

Comparing Box Plots



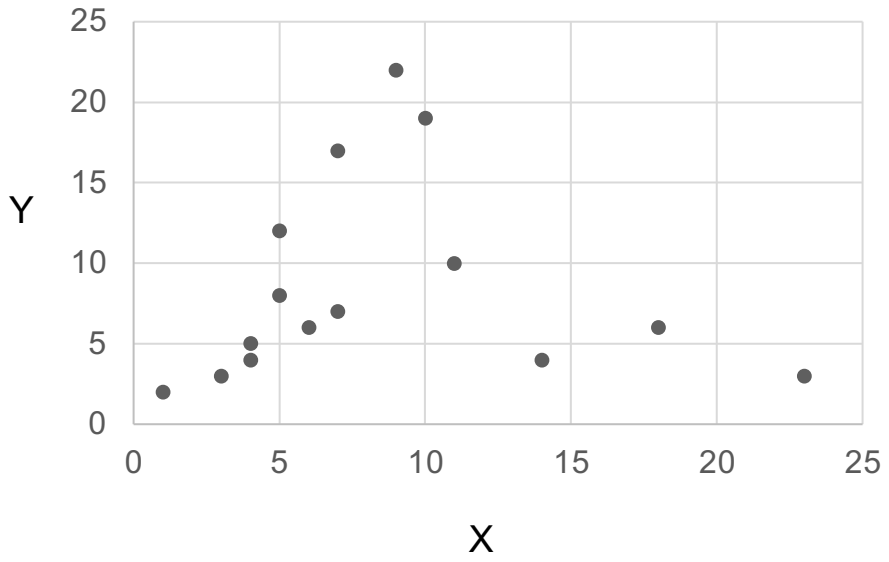
The distributions of boys scores on the test are negatively skewed, whilst the girls' score distribution is positively skewed. There are no outliers. The median score for boys is higher (M = 23) than for girls (M= 9.5). This IQR is smaller for boys (IQR = 10) than for girls (IQR = 12). The range of scores for boys and girls is equal (Range = 19).

Bivariate Data

Scatterplots

Numerical | Numerical

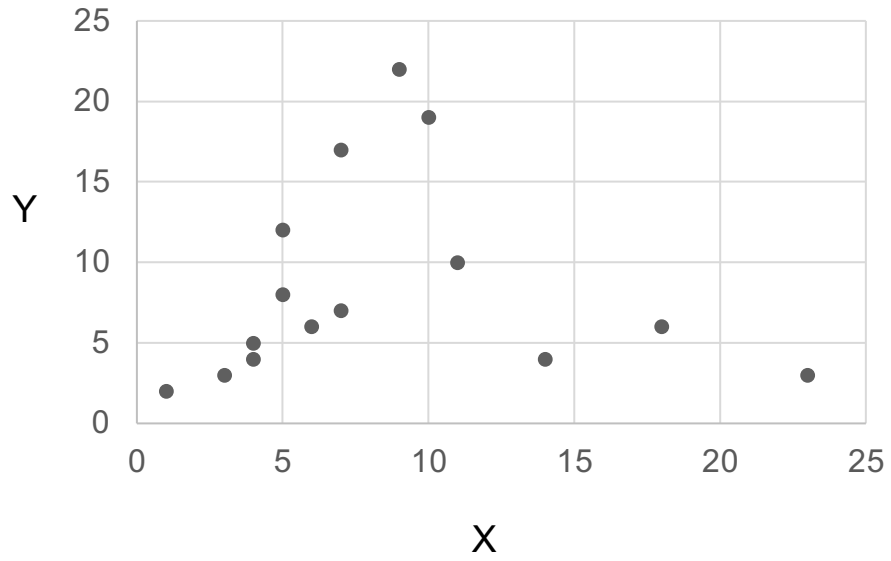
- Used HEAPS in the real world, super useful!
- Make sure you know how to plot these on your calculator.



Bivariate Data

Scatterplots

Numerical | Numerical



When describing scatterplots we MUST mention three things:

1. Strength
2. Direction
3. Form

Pearson's Correlation Coefficent

- Otherwise known as the *r* value
- Measures the ***strength of a linear relationship***
- We generally assume that a linear relationship is present (in some cases it isn't, but we'll get to that)
- Always find the value of r using your calculator, can't do it by hand.

Pearson Correlation Coefficient

- **Strong positive:** r = 0.75 to 0.99
- **Moderate positive:** r = 0.5 to 0.74
- **Weak positive:** r = 0.25 to 0.49
- **No association:** r = -0.24 to 0.24
- **Weak negative:** r = -0.25 to -0.49
- **Moderate negative:** r = -0.5 to -0.74
- **Strong negative:** r = -0.75 to -0.99

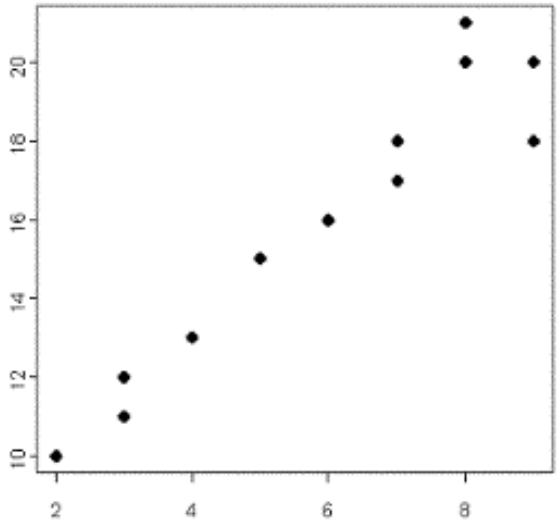
- Size of r value → STRENGTH of the association
- Sign in front of r value → DIRECTION of the association

Warning: Can only calculate r values for ***linear*** data sets

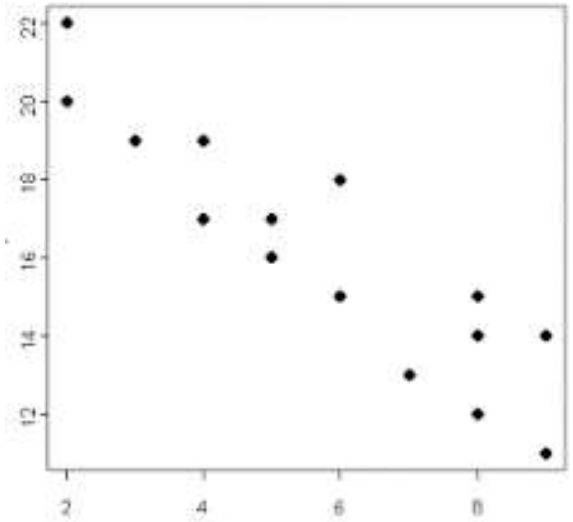
Bivariate Data

Direction

- Positive or negative
- A direction suggests that there is an association between two variables



Positive

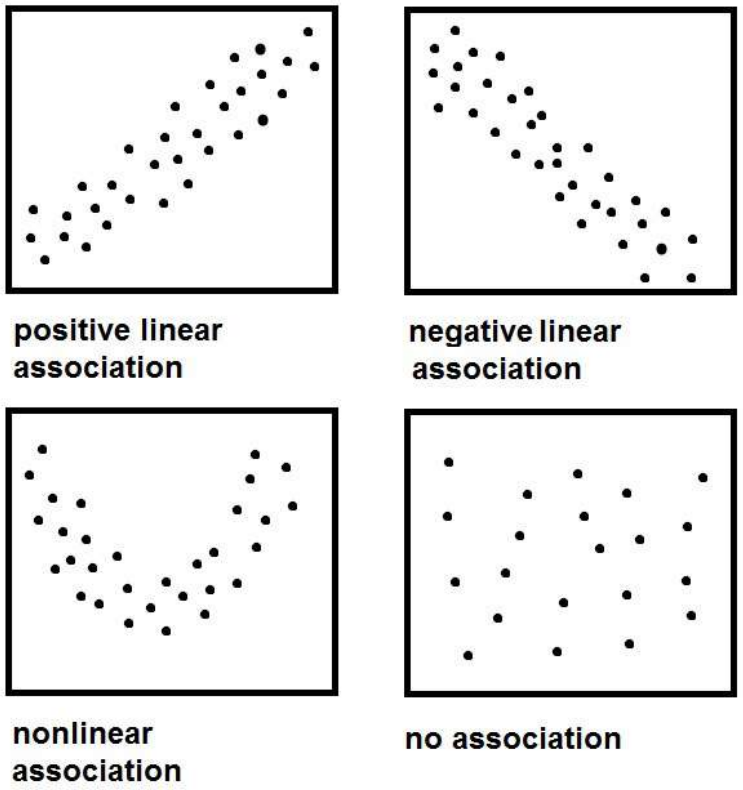


Negative

Bivariate Data

Form

- Linear, non-linear or no association



Linear: Data follows a relatively straight line.

Non linear: Data does not occur in a straight line patter, but does follow a curved pattern.

No association: Data points are randomly spread and do not appear to be associated.

Interpreting r

- If asked to interpret the value of the correlation coefficient use the following template sentences. Be sure to put these in your bound reference!

Linear, positive and strong It can be concluded that the **y variable** should increase as the **x variable increases**.	**Linear, positive and moderate** There is some evidence to suggest that the **y variable** should increase as the **x variable increases**.	**Linear, positive and weak** There is limited evidence to suggest that the **y variable** should increase as the **x variable increases**.
Linear, negative and strong It can be concluded that the **y variable** should decrease as the **x variable increases**.	**Linear, negative and moderate** There is some evidence to suggest that the **y variable** should decrease as the **x variable increases**.	**Linear, negative and weak** There is limited evidence to suggest that the **y variable** should decrease as the **x variable increases**.

Bivariate Data

Summary

Type of Data		Graph
Explanatory Variable	Response Variable	
Categorical	Categorical	Segmented Bar Chart or Two Way Frequency Table
Numerical	Categorical	Parallel Box Plots
Numerical	Categorical	Back to Back Stem and Leaf Plots
Numerical	Numerical	Scatterplots

Displaying Bivariate Data

- Strong positive: r = 0.75 to 0.99
- Moderate positive: r = 0.5 to 0.74
- Weak positive: r = 0.25 to 0.49
- No association: r = -0.24 to 0.24
- Weak negative: r = -0.25 to -0.49
- Moderate negative: r = -0.5 to -0.74
- Strong negative: r = -0.75 to -0.99

Describing bivariate distributions: SDF

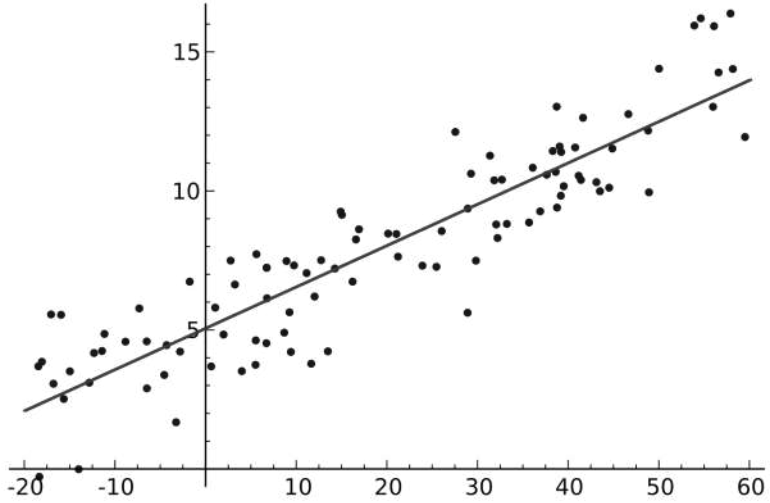
ATARNotes

3. Modelling Data

Modelling Data

- Bivariate data, and in particular bivariate data with two numerical variables, is extremely useful!

- This is because we can use it to construct *models*, mathematical equations that allow us to predict the values of data points we didn't even measure.

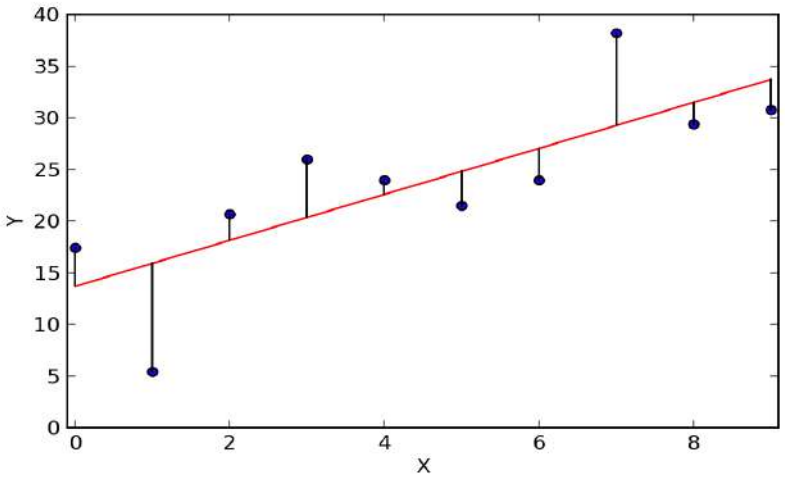


Least Square's Line of Best Fit

- How do we come up with a line of best fit?!?!

Few things:

- We take the *residual* = vertical distance between actual data point and line of best fit .
- We then make sure our line of best fit line minimises the sum of the squares of residuals
- Works best if there are ***no outliers***



Line of Best Fit

$$y = a + bx$$

- When we have a scatter-plot, we can find a *linear equation* that allows us to predict y from x.
- x is the explanatory variable
- y is the response variable
- **Must** ensure you are entering these variable in the correct order!
- Classic VCAA trick to give you the y variable before the x.

Forming the Line of Best Fit

- General form of a least squares line of best fit is:

$$y = a + bx$$

- Two scenarios:
  1. If you're given the raw data, use your CAS!
  2. If you're given statistics, use the following formulas!

Where:

- The slope/gradient of the line is $b = r \times \frac{s_y}{s_x}$
- The y intercept of the line is $a = \bar{y} - b\bar{x}$

And:

- r is the Pearson correlation coefficient
- s_y and s_x are the sample standard deviations of y and x respectively.
- $\bar{x}$ and $\bar{y}$ are the sample means of x and y

Modelling Data

Example

Height	160 cm	163 cm	165 cm	169 cm	174 cm	180 cm	185 cm	191 cm
Weight	60 kg	63 kg	70 kg	67 kg	72 kg	75 kg	71 kg	77 kg

Find the regression equation used to calculate height from weight

- Put data in the CAS (lists and spreadsheet)
- Place the height on the x axis and weight on y axis
- Find regression equation
- Weight = 58.022 + 1.63 x Height

Interpreting the Line of Best Fit

- We can interpret the regression line, $y = a + bx$ by saying:
  - The y intercept is a. This means the $y\ variable$ is $a\ units$ when the $x\ variable$ is zero $units$.
  - The slope is b. This means that the $y\ variable$ increases/decreases by $b\ units$ for every 1 $unit$ increase in the $x\ variable$.
- Use the word ‘increases’ when b is positive, and ‘decreases’ when b is negative.
- Replace everything in red to fit the context of the question.

Coefficient of Determination

- r^2 → Used where we can reasonably believe there is causation

- Tells us the extent to which x caused y

Explanation: The coefficient of determination tells us that r^2x100% of the variation in the response variable is explained by the variation in the explanatory variable

Warning: r^2 will always come out of the calculator positive.
We can tell if it is truly positive or negative by observing the scatterplot or gradient

Extra- and Interpolation

How reliable are these predictions?

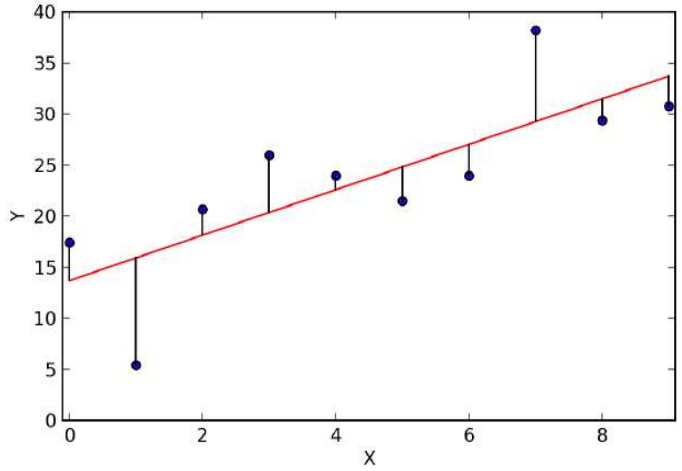
- "Assess the validity"

- Interpolation: The x value you are predicting from is within the data set (a fairly reliable prediction)

- Extrapolation: The x value you are predicting from is outside of the data set (an unreliable prediction)

Residuals

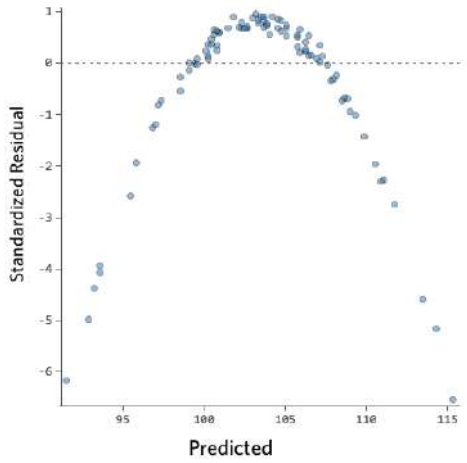
- How can we *mathematically* check if a scatterplot is linear or not?
- We plot a residual plot!
- To find the residual of a specific point, the equation is: $Residual = Actual\ y\ value\ - Predicted\ y\ value$

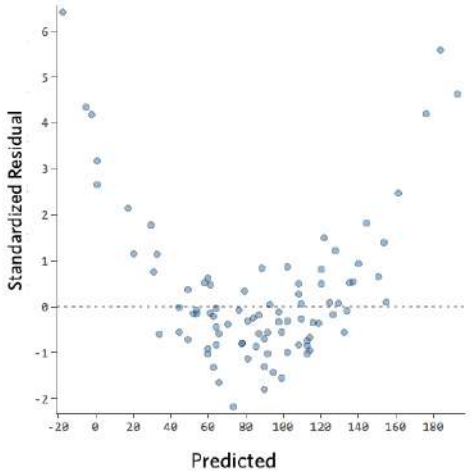


Modelling Data

RESIDUAL PLOTS

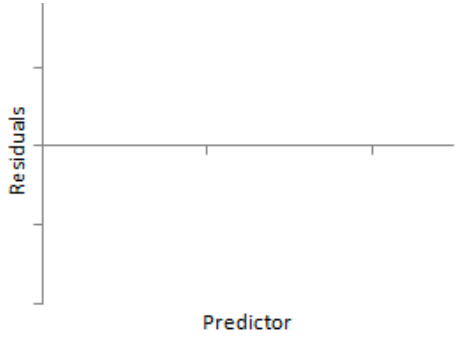
Non – linear



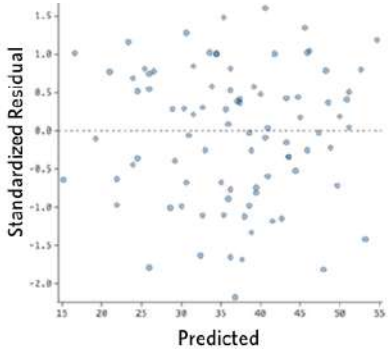


Clear Patterns

Linear

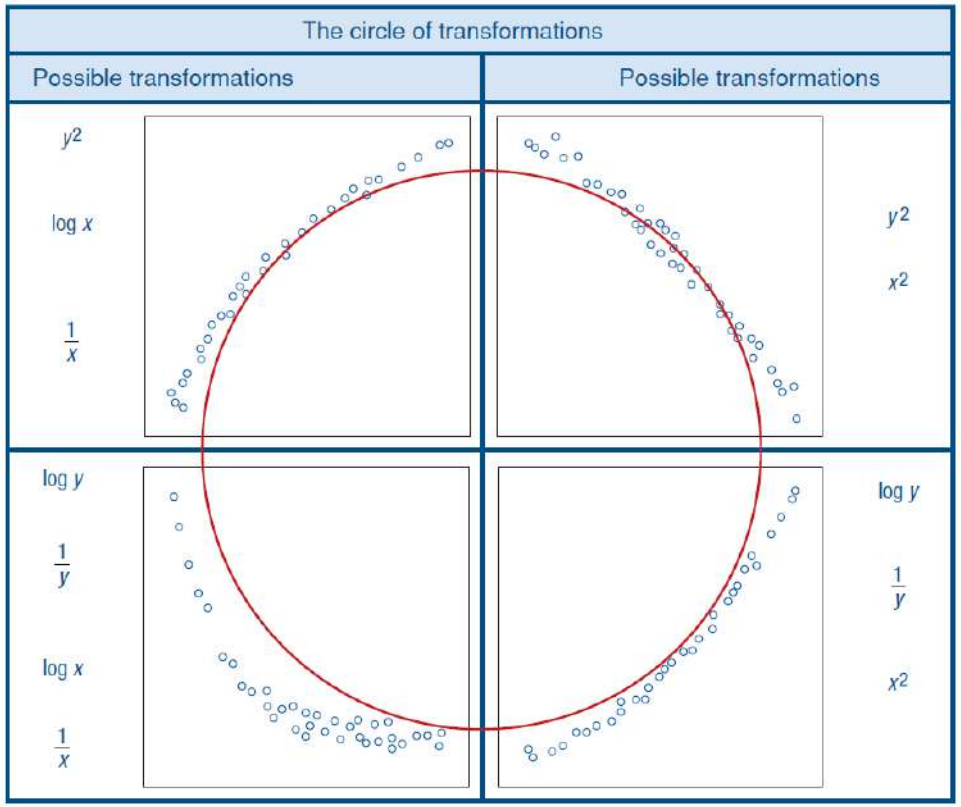


Every residual is zero



Residuals randomly scattered close to the x-axis

Data Transformation



Essential Further Mathematics Units 3 & 4, 4th Edition, pg 190.

Modelling Data

Example

Applying a x^2 transformation

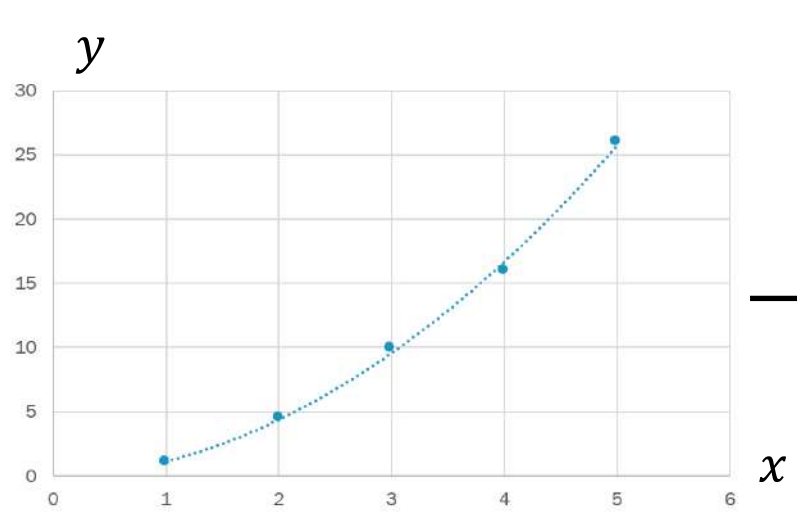
y	1	4.5	10	16	26
x	1	2	3	4	5
x^2	1	4	9	16	25

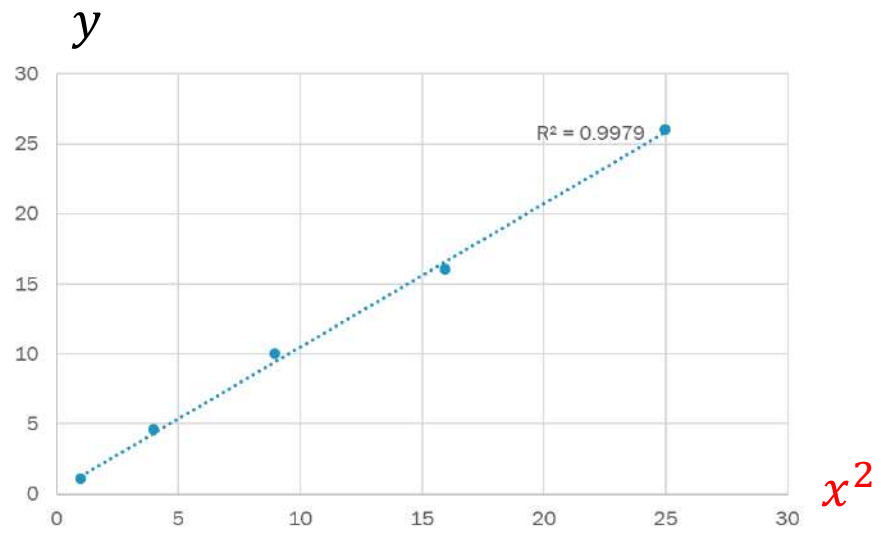
Therefore, we would graph and find the recurrence equation of

y	1	4.5	10	16	26
X^2	1	4	9	16	25

Modelling Data

Example

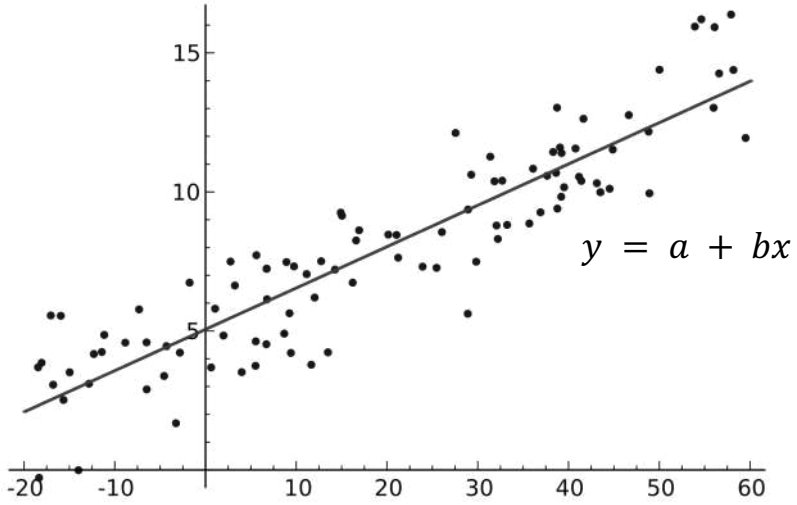




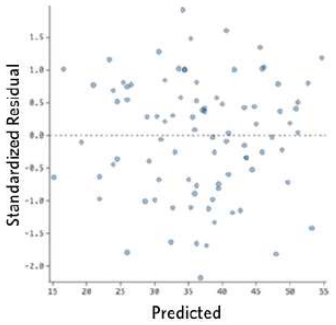
Pre transformation, non –linear

Post – transformation
→ Data has been linearised

Modelling Data

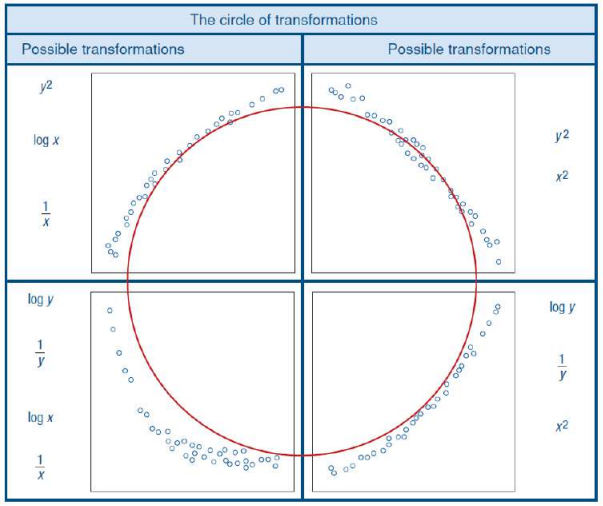


Interpreting and calculating lines of best fit



Residual plots

Summary



Essential Further Mathematics Units 3 & 4, 4th Edition, pg 190.

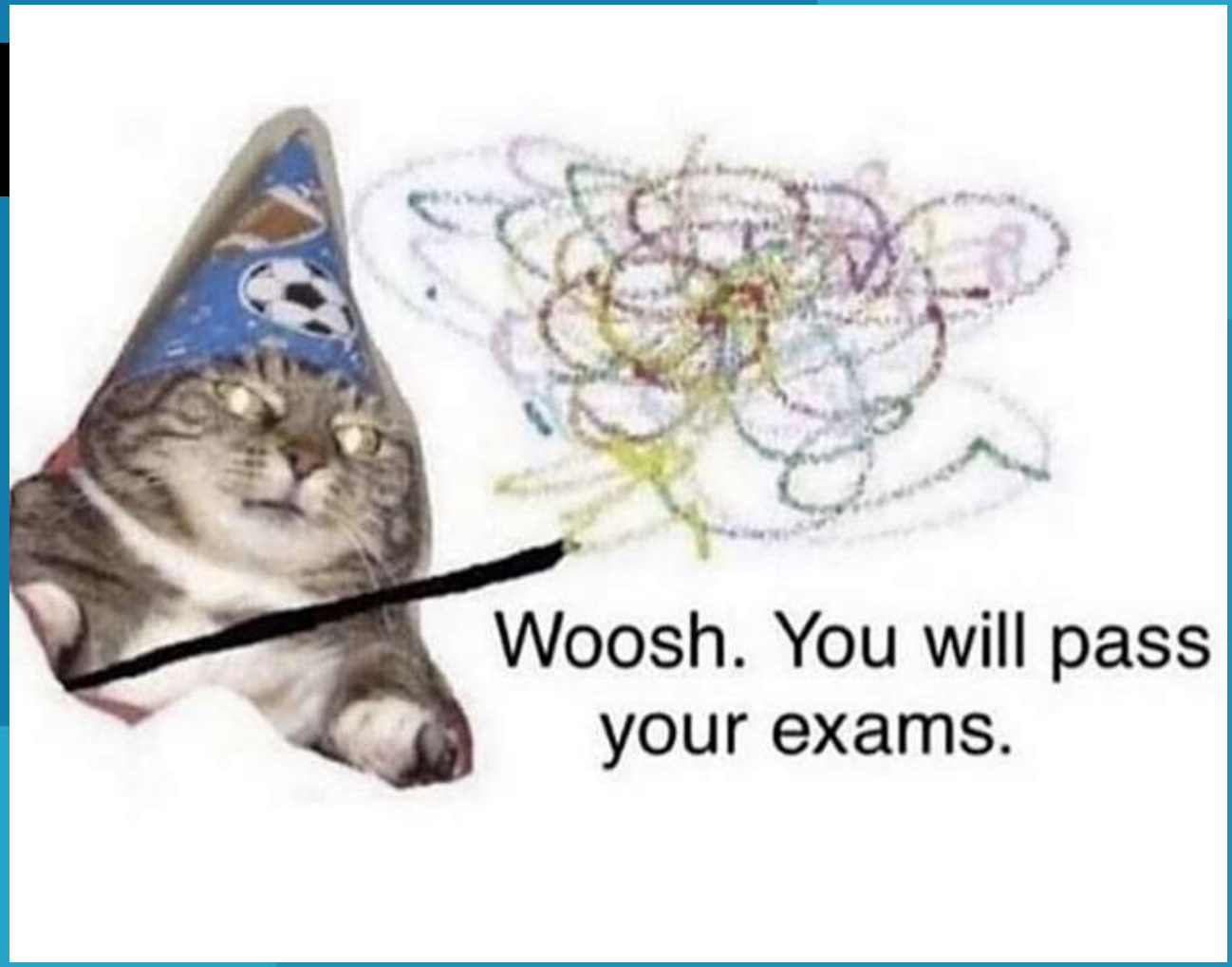
Transformations

r^2

Correlation coefficient

ATARNotes

GOOD LUCK <3



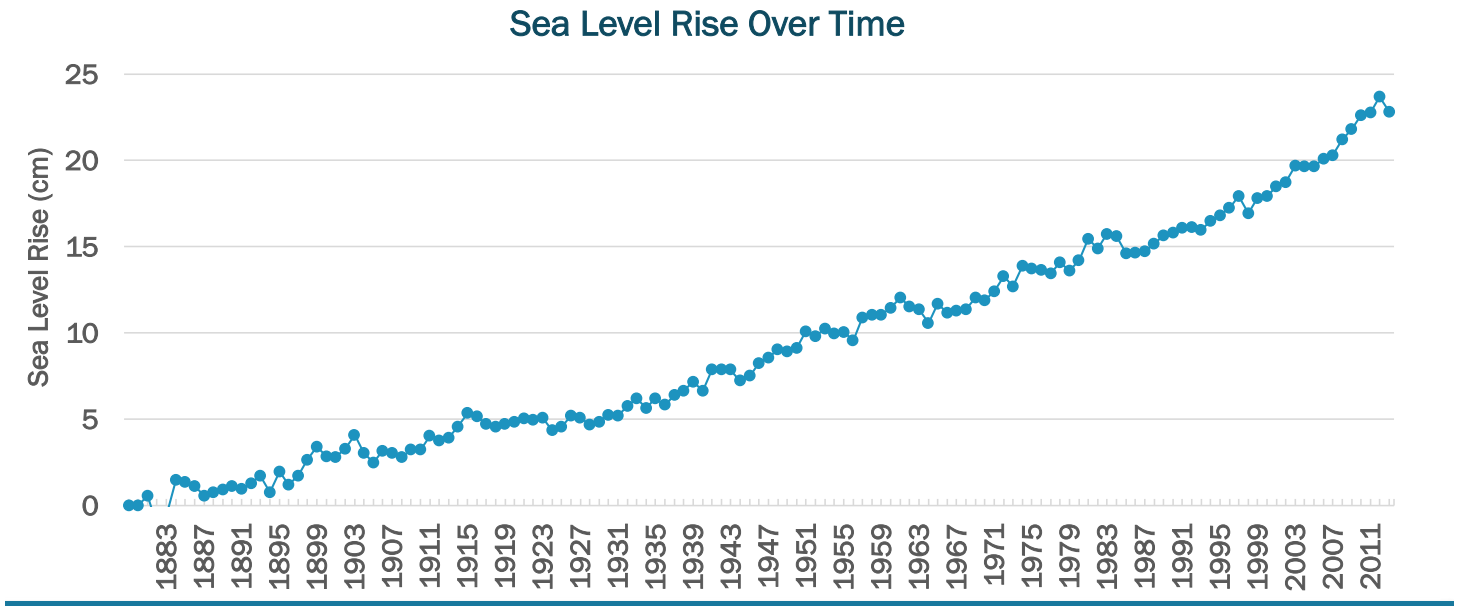
ATARNotes

4. Time Series
(Extra content completing U3 AOS1, will not be covered in recording)

Time Series

The *same* as a regular scatterplot, except:

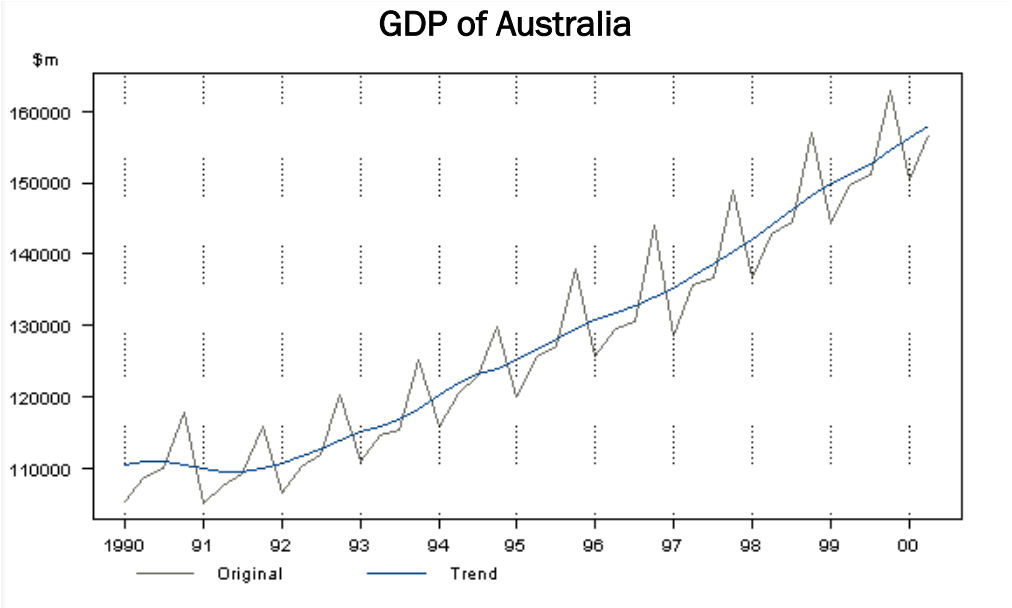
- Our explanatory (x) variable is time
- We connect the data points with lines



Trend

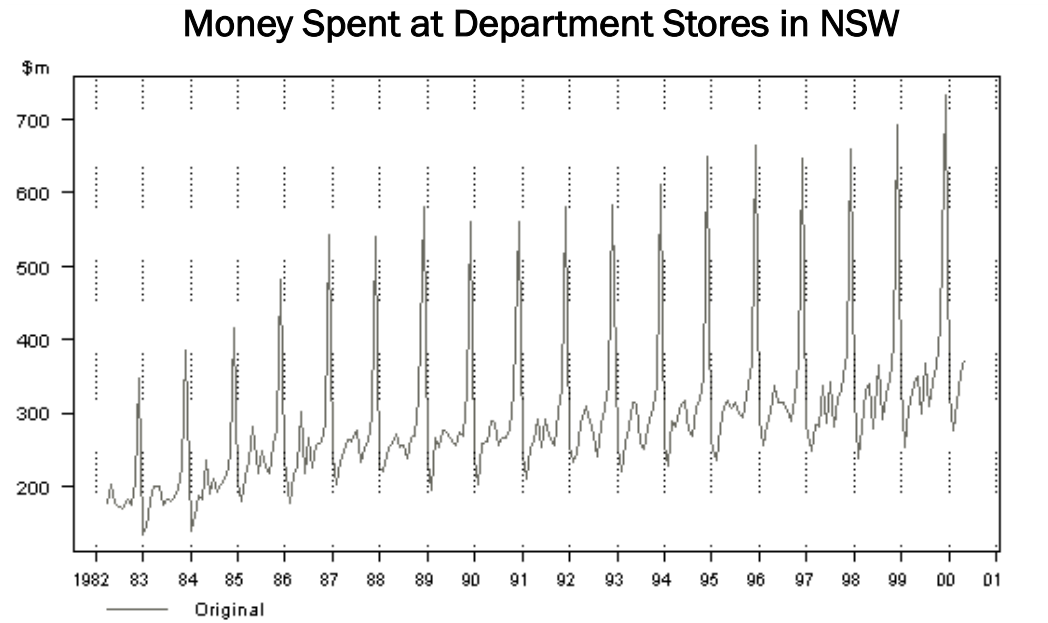
- Describes what is happening in the ***long term***

- Increasing Trend: Present where there is a positive slope
- Decreasing Trend: Present where there is a negative slope



Seasonality

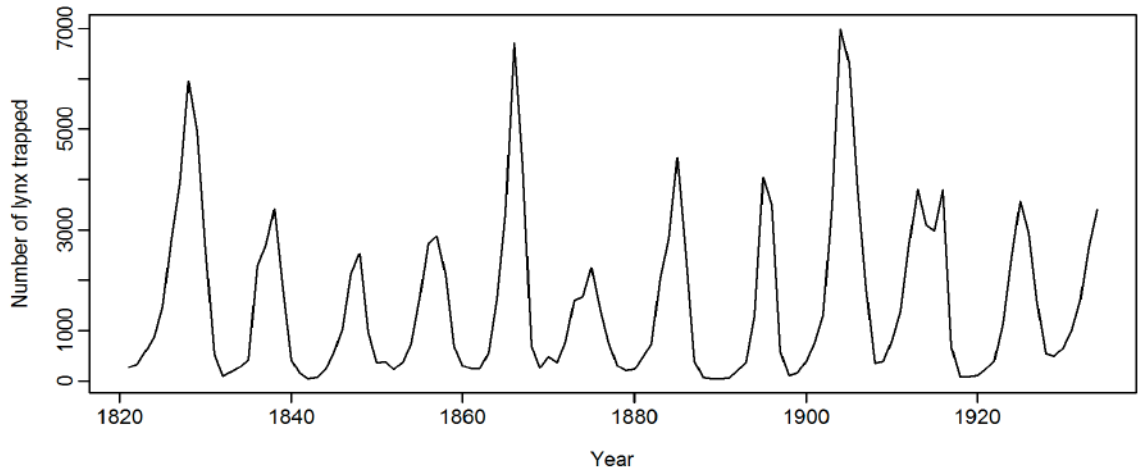
- Peaks/troughs at regular intervals related to the calendar (usually seasons of the year, but could be weekly, monthly etc.)
- Typically similar in size
- Some consistency to the peaks/troughs



http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

Cycles

- Long terms variations that are not seasonal
- Cycles do **NOT** exist within a year, they are only present in time series extending more than a year
- Seasonality can exist within cycles



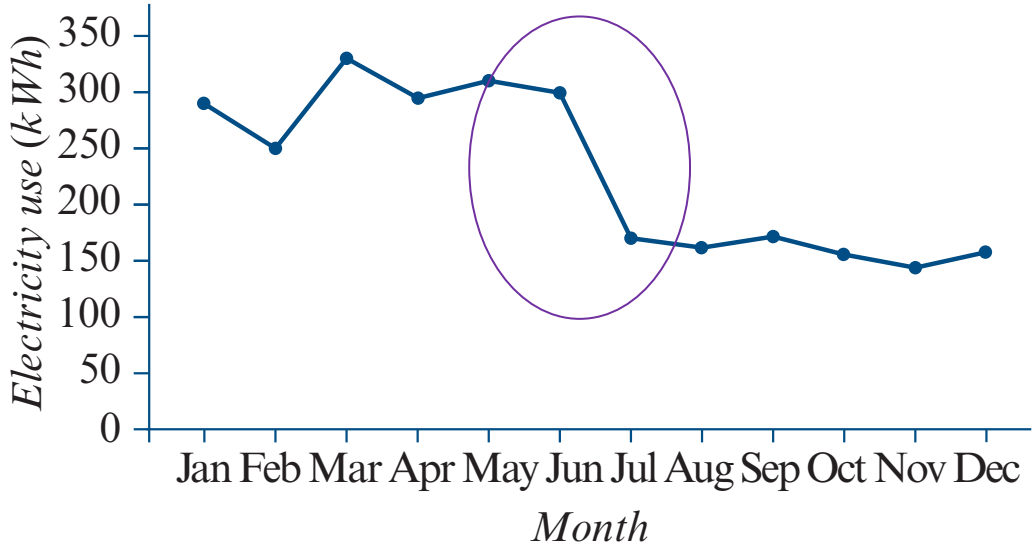
Irregular Fluctations

- These are present in basically every times series we look at
- Any data point that cannot be attributed to cycles, seasonality, trends or structural change are classified as irregular
- Basically, if ever there is a data point that isn't perfectly in place, we say there are irregular fluctuations (this was every time series I came across)
- If you've got no idea just guess this lol

Structural Change

- Where there is a sudden change in the established pattern of a time series plot

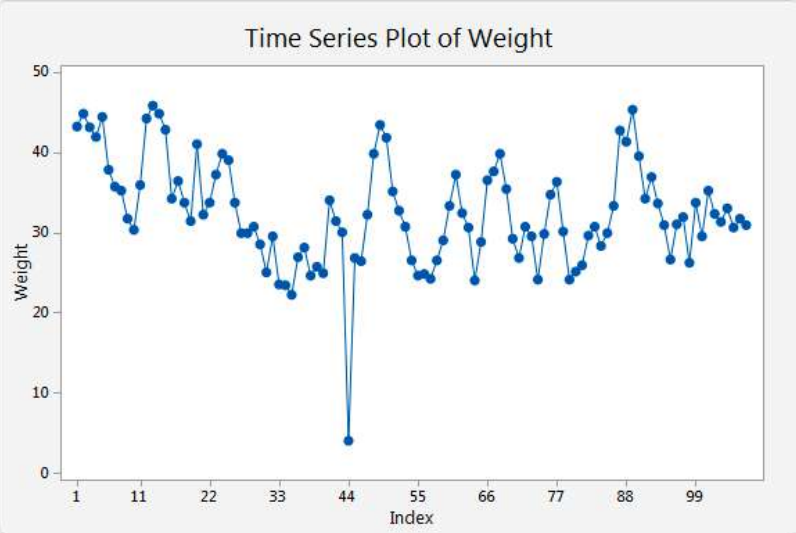
- Must be a marked change that is then continued in subsequent data



Outliers

- Individual data points that stand out from the general body of data

- Generally caused by a one off event, common example is a financial crisis



Smoothing Time Series

- Most of the time, time series plots look pretty messy, and this makes them hard to read.
- To make the trends on the series a little easier to identify, we use a process called *smoothing*.
- Two methods:

1. Moving-mean smoothing (numerical)
2. Moving-median smoothing (graphical)

Moving Mean Smoothing

- Dilutes the effect of large fluctuations
- Basically takes into account the surrounding data to each point to give a clearer trend flow

Time	y	3-moving mean	Smoothed y
1	7		
2	13	(7+13+6)/3	8.67
3	6	(13+6+14)/3	11
4	14	(6+14+6.5)/3	8.83
5	6.5		

Moving Mean Smoothing

- Slightly more complicated for even number smoothing

Time	y	4-moving mean	Centring	Smoothed y
1	7			
2	13			
		7 + 13 + 6 + 14 = 40 / 4 = 10		
3	6		(10 + 9.88)/2 = 9.94	9.94
		13 + 6 + 14 + 6.5 = 39.5 / 4 = 9.88		
4	14			
5	6.5			

Moving Median Smoothing

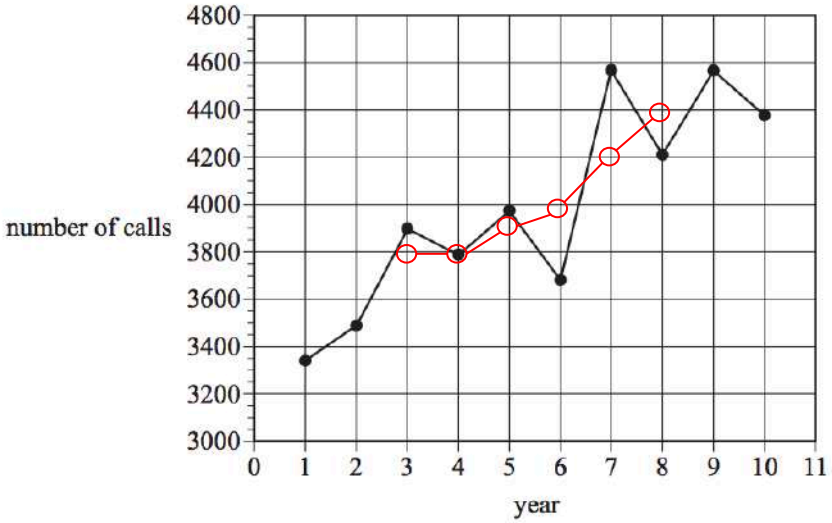
- Uses the graphical representation of data points to find a smoothed line

Tip: Double check your answers! Math by sight leaves you open to errors

Moving Median Smoothing

Question 12

The time series plot below charts the number of calls per year to a computer help centre over a 10-year period.



Using five-median smoothing, the smoothed number of calls in year 6 was closest to

A. 3500
B. 3700
C. 3800
D. 4000
E. 4200

2015 VCAA Exam 1

Seasonal Indices

- Another way that we can help make data more easily readable is through the use of *seasonal indices*.
- Sometimes, we want to compare and make regression lines (say perhaps, sales figures for a business) for time series data.
- However, seasonality can make it hard to accurately get a linear relationship.
- To overcome the effects of seasonality, we can ***deseasonalise*** our data using seasonal indices.

Seasonal Indicies

$$seasonal\ index = \frac{value\ of\ season}{seasonal\ average}$$

- If we add all of the seasonal indices in a data set we are given the number of seasons (generally this is 4 as most data looks at a whole year)

- Note: season can mean various things:
  - Month
  - Quarter
  - Weather Seasons

Interpreting Seasonal Indicies

Interpreting:

A seasonal index of 1.3 during summer tells us that figures for summer are 30% **above** average

A seasonal index of 0.87 during winter tells us that figures for winter are 13% **below** average

Correcting with Seasonal Indices

Correcting:

To correct for seasonality, our formula is $\frac{1}{seasonal\ index}$

Warning: People always screw this up, make sure you get it!

Eg. January's seasonal index is 0.8

To correct for seasonality, we should increase the figures for January by 25% because 1/0.8=1.25 (125%)

Deseasonalisation

- Seasonal variation complicates regression, so we deseasonalise data before fitting a line

- Predictions **must** however take into account seasonal variation

- So, we **reseasonalise** data predicted from our equation

Deseasonalised Values

- To deseasonalise:
  - $Deseasonalised\ value = \frac{actual\ value}{seasonal\ index}$
- To reseasonalise predicted data:
  - $Actual\ value = deseasonalised\ value \times seasonal\ index$

Making Forecasts

1. Calculate seasonal indices
2. Deseasonalise the response variable
3. Fit a line to the deseasonalised data
4. Make predictions by substituting in values for time to get a deseasonalised prediction
5. Reseasonalise this prediction to get the actual prediction

Example

	Summer ’11	Autumn ’11	Winter ‘11	Spring ‘11
Sales	120	93	65	108
SI				

1. Calculate seasonal indices

120 + 93 + 65 + 108 = 386
386 / 4 = 96.5 (seasonal average)

Summer ’11 SI = 120 / 96.5 = 1.24
Autumn ‘11 SI = 93 / 96.5 = 0.96

(checking all answers add to 4)

Example

	Summer ’11	Autumn ’11	Winter ‘11	Spring ‘11
Sales	120	93	65	108
SI				

2. Deseasonalise the response variable

120/1.24 = 96.77
93/0.96 = 96.88
65/0.67 = 97.01
108/1.12 = 96.43

Example

	1	2	3	4
Sales	120	93	65	108
SI	1.24	0.96	0.67	1.12
Deseas Sales	96.77	96.88	97.01	96.43

3. Fit a line to the deseasonalised data

- To fit a line, you must number each point on the x axis
- Calculate this line the same way you would a line of best fit

$$Deseasonalised\ sales = 97 - 0.089 \times quarter\ number$$

EXAMPLE

	1	2	3	4
Sales	120	93	65	108
SI	1.24	0.96	0.67	1.12
Deseas Sales	96.77	96.88	97.01	96.43

4. Make predictions using regression line

Predict the deseasonalised sales for Spring '12:

$$97 - 0.089 \times 8 = 96.23$$

EXAMPLE

	1	2	3	4
Sales	120	93	65	108
SI	1.24	0.96	0.67	1.12
Deseas Sales	96.77	96.88	97.01	96.43

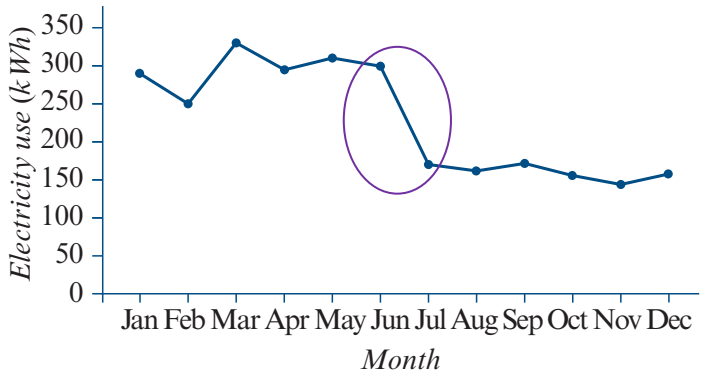
5. Reseasonalise to get actual prediction

$$96.23 \times 1.12 = 107.78$$

$$(deseasonalised\ value \times seasonal\ index = Actual\ value)$$

Time Series

Summary



- Trends
- Seasonality
- Cycles
- Irregular fluctuations
- Structural changes

Time	y	3-moving mean	Smoothed y
1	7		
2	13	(7+13+6)/3	8.67
3	6	(13+6+14)/3	11
4	14	(6+14+6.5)/3	8.83
5	6.5		

Numerical and graphical smoothing

$$\text{seasonal index} = \frac{\text{value of season}}{\text{seasonal average}}$$

$$AV = DV \times SI$$

$$DV = \frac{AV}{SI}$$

Seasonal indices and deseasonalisation

Forecasting